UNIVERSITY OF TRENTO - Italy

International PhD Program in Biomolecular Sciences

XXVII Cycle

A comparative analysis of the metabolomes of different berry tissues between *Vitis vinifera* and wild American *Vitis* species, supported by a computer-assisted identification strategy

Tutor

Dr. Fulvio Mattivi

Department of Food Quality and Nutrition, Fondazione Edmund Mach

Advisor

Prof. Vladimir Shulaev

Department of Biological Sciences, University of North Texas

Ph.D. Thesis of

Luca Narduzzi

Department of Food Quality and Nutrition

Fondazione Edmund Mach

Academic Year 2013-2014

Declaration

I, Luca Narduzzi, confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Acknowledgements

An endless list of people contributed to the success of my Ph.D. experience, and name them one by one is almost impossible, unless each of them deserves it.

The first thank goes to my family that even if they do not know how, they made it possible. They silently supported my discoveries, which became more and more un-understandable to them. However, their curiosity and support never step back.

I would like to thank my tutor, Dr. Fulvio Mattivi, to be a very enlightening guidance; he let me follow my own inspiration and let me develop my own ideas, without trapping me in research dogmas, which are the results of someone else's thinking.

I am in debit with Dr. Pietro Franceschi, who helped me as if I was his own student, with endless patience. I ought to pay something back, sooner or later.

Big thanks go to Dr. Panagiotis Arapitsas to act as the devil part (Ron docet) and for continuously pushing me on being more concrete. Big thanks go also to Dr. Vladimir Shulaev, which hosted me in his lab for several months and he pushed me through specific research paths.

Big big respects go to Manoj Ghaste, my Indian counterpart, who shared with me the joys and the dramas of this trip. In any journey, traveling with a friend is hundred times better than travel alone. Thanks Manoj.

In general, I need to thank the people that spent time with me: all the lab people, the ones that shared just few days, and the ones that spent with me the whole Ph.D. period, or part of it. Company and friendship are priceless gifts that many people decided to give me and I am very grateful with all of you guys. I also want to thank my "soccer" team, which converted me in a "bomber" and helped me in scoring the best goals of my life.

In the last year, I met Francesca, who accompanied me through the hardest part of this trip. It did not come out as I expected and your support has been necessary to go through all of it. If we are together, failing is not so bad, after all.

During my whole academic carrier, I met people that influenced my path and eventually drove me back on the right way. They do not know it, but I would like to thank them for doing so. Thanks go to: Dr. Fabiana Fabbretti, Prof. Giorgio Prantera and Dr. Silvia Volpi to let me understand that future is always in our hands. Samuela Palombieri, that pushed me though the Erasmus experience, which has been the key point of my whole academic career. To Jan Stanstrup, this showed me that nothing is impossible (in metabolomics).

I would like to thank Cristina Mattei; she has been a torch when my light was off, bearing all my needs and listening to all my doubts. She has been continuously challenging me to go towards the next level, like in a videogame, but playing with real lives.

Last, but first in importance, I must thank Dr. Maria Pilar Lopez Gresa, which first infused my mind with her curiosity for sciences, and later, she believed in my possibilities when no one else did.

Abstract

Grape (*Vitis vinifera L.)* is among the most cultivated plants in the world. Its origin traces back to the Neolithic era, when the first human communities started to domesticate wild *Vitis sylvestris L.* grapes to produce wines. Domestication modified *Vitis vinifera* to assume characteristics imparted from the humans, selecting desired traits (e.g. specific aromas), and excluding the undesired ones. This process made this species very different from all the other wild grape species existing around the world, including its progenitor, *Vitis sylvestris*.

Metabolomics is a field of the sciences that comparatively studies the whole metabolite set of two (or more) groups of samples, to point out the chemical diversity and infer on the variability in the metabolic pathways between the groups. Crude metabolomics observation can be often used for hypotheses generation, which need to be confirmed by further experiments. In my case, starting from the grape metabolome project (Mattivi et al. unpublished data), I had the opportunity to put hands on a huge dataset built on the berries of over 100 *Vitis vinifera* grape varieties, tens of grape interspecific hybrids and few wild grape species analyzed per four years; all included in a single experiment. Starting from this data handling, I designed specific experiments to confirm the hypotheses generated from the observation of the data, to improve compound identification, to give statistical meaning to the differences, to localize the metabolites in the berries and extrapolate further information on the variability existing among the grape genus.

The hypotheses formulated were two: 1) several glyco-conjugated volatiles can be detected, identified and quantified in untargeted reverses-phase liquid chromatography-mass spectrometry; 2) The chemical difference between *Vitis vinifera* and wild grape berries is wider than reported in literature. Furthermore, handling a huge dataset of chemical standards injected under the same conditions of the sample set, I also formulated a third hypothesis: 3) metabolites with similar chemical structures are more likely to generate similar signals in LC-MS, therefore the combined use of the signals can predict the more likely chemical structure of unknown markers.

In the first study (chapter 5), the signals putatively corresponding to glycoconjugated volatiles have been first enclosed in a specific portion of the temporal and spectrometric space of the LC-HRMS chromatograms, then they have been subjected to MS/MS analysis and lastly their putative identity have been confirmed through peak intensity correlation between the signals measured in LC-HRMS and GC-MS.

In the second study (chapter 6), a multivariate regression model has been built between LC-HRMS signals and the substructures composing the molecular structure of the compounds and its accuracy and efficacy in substructure prediction have been demonstrated.

In the third study (chapter 7), I comparatively studied some wild grapes versus some *Vitis vinifera* varieties separating the basic components of the grape berry (skin, flesh and seeds), with the aim to identify all the detected metabolites that differentiate the two groups, which determine a difference in quality between the wild versus domesticated grapes, especially regarding wine production.

Index

1. General Introduction

1.1 The genus *Vitis* and the *Vitis vinifera*

1.1.1 The *Vitis* genus

The genus *Vitis* is part of the angiosperms of order *Vitales*, family *Vitaceae. Vitis* genus consists of about 60 eco-species, most of which are inter-fertile and originate from the northern hemisphere of the world. The origin of the *Vitis* genus dates back about 60 million years ago in the Paleocene epoch. All the species of this genus are arboreal plants anchored to the ground through the roots that collect nutrients for the plant growth. The shoots sprout from the woody trunk, and they have nodes from where new leaves and flowers can form. Generally, *Vitis* plants have tendrils, green elongations of the trunk that are able to cling to handholds, helping the plant to achieve a vertical position. Bushy *Vitis* exist, but are a minority.

Vitis leaves are generally peculiar, with a very expanse area; their shape depends mostly upon the species and on the variety. In general, *Vitis* leaves connect to the trunk by a petiole long 5 to 10 centimeters; from the connection between the petiole and the leaf, start five ribs that carry the nutrients through the five lobes of the leaf. The leaves perform C3 carbon fixation, therefore are not resistant to extreme drought.

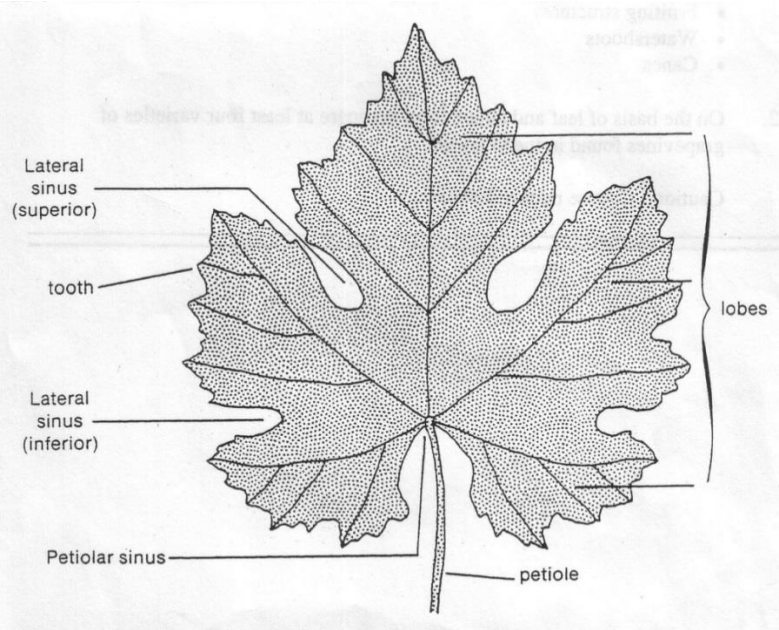


Image 1: example of grape leaf. In the image we can distinguish at the bottom the petiole sinus where 5 main ribs spread across the 5 lobes. The shape of the lobes, tooth and sinus are different per each variety and each different specie. These characteristics have been studied in a field called Ampelography to distinguish the different grapes.

Most of the *Vitis* species are wind-pollinated; they have hermaphrodite flowers grouped in inflorescences. Flowering happens when temperature falls around 15 to 20 centigrade (around May in the northern hemisphere and November in the Southern). When pollinated, each flower turn into grape berry, so the inflorescence becomes a cluster of grapes. The shape, color and number of the berries are strictly depending on the species and the variety, and it can be from tens to hundreds per cluster, in a close or open bunch. In nature, red and black grape berries are the most common ones, but white berries exist, especially in the domesticated grapes. The amount of seeds in the berry varies between one and four depending on variety, species and berry shape.

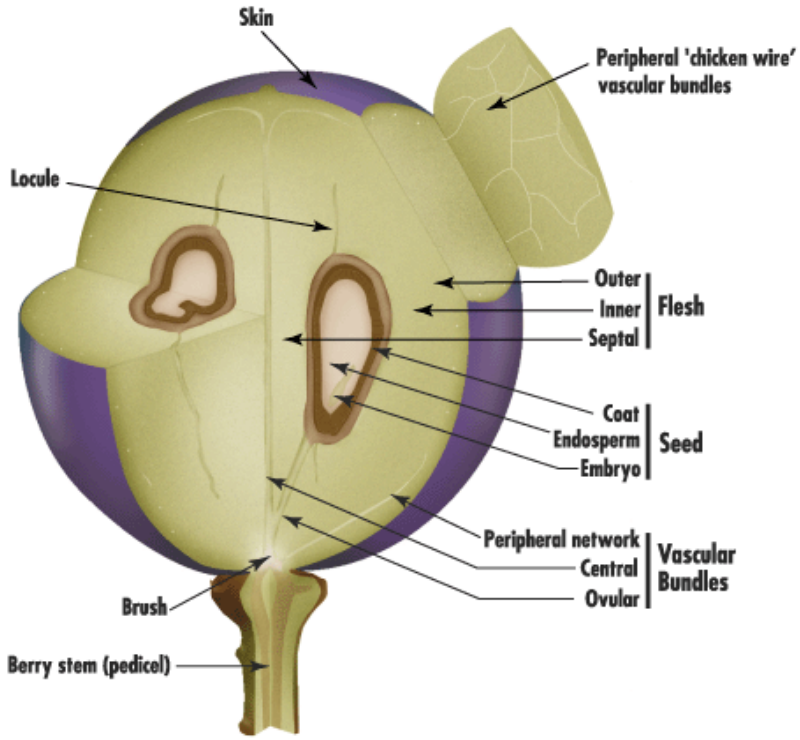


Image 2: An illustrative representation of the grape berry tissues. The three main tissues (Skin, flesh and seeds) are consistent of many different sub-tissues as shown in the image. Even if not directly represented in the image, also the skin is constituted of two different tissues, an external layer, and an internal one. The distribution of the compounds contained in the berry is very different both between tissues, and between the sub-tissues.

The shape of the berry is commonly rounded, even though oblong grape berries exist. It is possible to compare the size measuring the diameter. Generally domesticated and selected grape have a diameter range of 10 to 25 millimeters, while wild grape range across 3 to 15 millimeters. Being generally a sphere, the ratio between the amounts of skin/flesh changes according to the berry size. The skin might be very thick, especially in wild *Vitis*; furthermore, the amount and size of the seeds is variable at the expense of the amount of flesh. The ratio of the amount of skin/flesh/seeds is a peculiarity of the variety and the specie studied. The variability existing between the berries is very

wide within the *Vitis* genus; it is beyond this thesis discuss and classify the numerous possibilities that we might encounter in all the germplasm. As mentioned earlier, a field of botany, Ampelography, since the XIX century systematically classify grapes according to their leaves shape and color (mostly). For the reader, interested on knowing more about this field, I report four main works:

1. *Ampelography,* published by Pierre Viala and Victor Vermorel in 1910 (in French), which is a seven books encyclopedia describing all the varieties and species known at that time, with hand drawings of the grapes (the hand drawings are published also in this thesis).
2. *"Origin and classification of cultivated grape"* in USSR ampelography by prof. Alexander Mikhailovich Negrul who studied the ampelographic differences of the *Vitis vinifera* varieties in the Asian territories. Through ampelographic classification, he was able to classify *Vitis vinifera* in three main group ("proles"): A) "Proles Pontica" genotypes diffused in Georgia and in the Near-East. B) "Proles Occidentalis" genotypes spread in the western European countries (Italy, France, Spain and Germany). C) "Proles Orientalis" genotypes spread in central Asia, Afghanistan and Iran.
3. *Ampelographie practique* a book published by Pierre Galet in 1952, and translated in English in two different editions in 1979 and in 2000. The peculiarity of this book is that Galet invented a very sharp method to distinguish the different grape varieties based on multiple factors: "shape and contours of the leaves, the characteristics of growing shoots, shoot tips, petioles, the sex of the flowers, the shape of the grape clusters and the color, size and pips of the grapes themselves".
4. The work of Chitwood et al. (2014) tried to individuate the genetic bases of the leaf shape of 1200 different grape varieties.

In the last decade, the application of molecular biology and genetic markers enhanced the grape genotyping and now multiple methods based on "Single Sequence Repeats" (SSR) and "Single Nucleotide Polymorphisms" SNPs are used to classify and characterize *Vitis* germplasm collected in various research institutes in the world. A deeper description of the classification given by this approach is demanded to the literature, for example Lamboy (1998), This et al. (2004), Myles et al., (2011), and Emanuelli et al. (2013). The latter study is based on the germplasm maintained at the experimental fields of the Fondazione Edmund Mach, the institution where I developed my thesis, so all the classification of the grape materials used in this thesis is from the work of Emanuelli et al. (2013). A repository of all the *Vitis* characterized, product of a large collaborative European project,

exist on the website of the "*Vitis* International Variety Catalogue", www.vivc.de, where the classification of Emanuelli et al. (2013) has been uploaded.

1.1.2 *Vitis vinifera* origin, distribution and cultivation

Vitis vinifera is the domesticated progeny of *Vitis sylvestris*, a species originated probably around 45 million years ago in the Caucasus region, that naturally widespread in the entire near east and all the Mediterranean area. The domestication started around 8000 years ago, when humans started farming cereals and arboreal plants (like grape); jars containing trace of wines were found in many ruins of Neolithic cities and villages. Sumerian, Egyptian, Hittites, Persian and other old civilities have been reported to produce and consume wine, while the oldest wine press has been located in Armenia (southern Caucasus). In a recent paper, Myles et al. (2011) could establish the start of the domestication through a genetic analysis, confirming that all the *Vitis vinifera* are genetically closer to the *Vitis sylvestris* originating from the Caucasus than from the ones from western Mediterranean area. Nevertheless, *Vitis vinifera* also spread across the Mediterranean, and the further influence that western *Sylvestris* might have given to the domesticated *viniferas* is still under consideration (Emanuelli et al. 2013). A few Italian cultivars (Lambrusco di Sorbara, Enantio) have been classified as product of the domestication of autochthonous *Vitis Sylvestris* (Emanuelli et al. 2013).

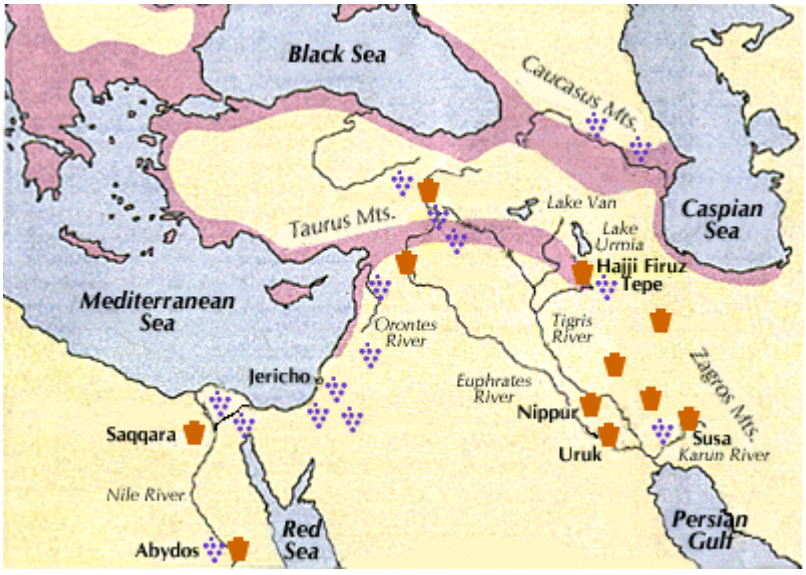


Image 3: The origin of *Vitis Sylvestris* is located in the southern area of the Caucasus region. In Southern Caucasus was found the first winery of the world. In this image, the first wineries are displayed, with an indication of the zones of diffusion of *Vitis Sylvestris.* It is probably in this area where domestication of *Vitis Sylvestris*started, producing the first *Vitis viniferas.*

During Roman Empire, wine production and consumption spread all around the Mediterranean Sea and Europe, and since then, it is still among the most cultivated arboreal plant in the world.

Domestication and diffusion of selected grapes (varieties) has been achieved through vegetative propagation, which has the advantage to control the genetic diversity and assure the quality of the products, but although, it is also responsible for the lack of natural selection and the weakening of natural defenses of the grapes against their natural pathogens. In the XIX century, grapes imported from the new world (Americas), brought to Europe also new pathogens, which almost destroyed *Vitis vinifera* cultivation. Indeed the main cause of the European grape blight was *Daktulosphaira vitifoliae*, commonly called "Phylloxera", an aphid that attacks the roots of the *Vitis vinifera*, to depose eggs, infecting it with a poisonous liquid that eventually kills the vines.

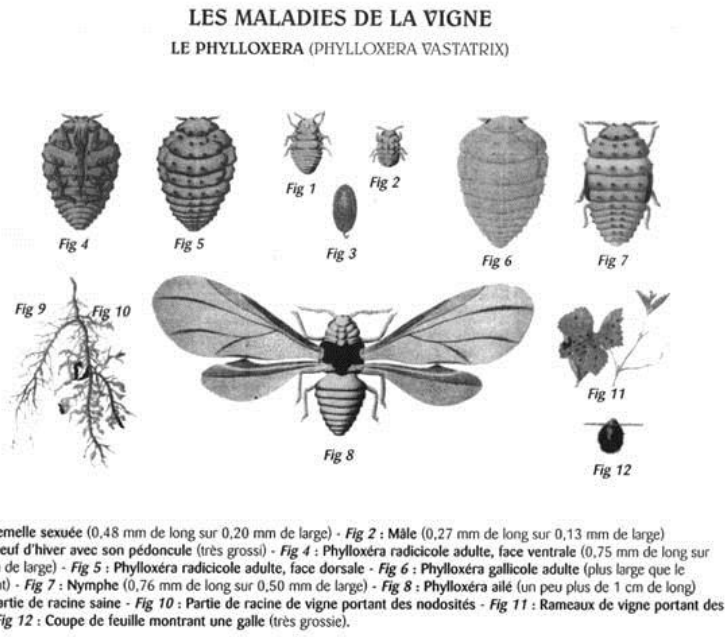


Fig 1 : **Femelle sexuée** (0,48 mm de long sur 0,20 mm de large) - *Fig 2* : **Mâle** (0,27 mm de long sur 0,13 mm de large) *Fig 3* : **Oeuf d'hiver avec son pédoncule** (très grossi) - *Fig 4* : **Phylloxéra radicicole adulte, face ventrale** (0,75 mm de long sur 0,50 mm de large) - *Fig 5* : **Phylloxéra radicicole adulte, face dorsale** - *Fig 6* : **Phylloxéra gallicole adulte** (plus large que le précédent) - *Fig 7* : **Nymphe** (0,76 mm de long sur 0,50 mm de large) - *Fig 8* : **Phylloxéra ailé** (un peu plus de 1 cm de long) *Fig 9* : **Partie de racine saine** - *Fig 10* : **Partie de racine de vigne portant des nodosités** - *Fig 11* : **Rameaux de vigne portant des galles** - *Fig 12* : **Coupe de feuille montrant une galle** (très grossie).

Image 4: In this image, the description of the life cycle of the Phylloxera from an old French biology book. As you may notice, the official name was different, and was indicating the "devastating" effect of the Phylloxera.

Between 1860 to 1880, "Phylloxera" and other minor parasites caused the destruction of over two thirds of the European vineyards, which brought to a drop in *Vitis vinifera* variability. In the last decades of the XIX century, grape producers, to defend *Vitis vinifera* from the pests, needed to graft all the vines on American rootstocks that proved to be resistant to the pathogens. A second solution was to create crosses between *Vitis vinifera* x American *Vitis* that were chosen selectively to resist the pests and produce quality wines. The description of all the existing American species, rootstocks and the resistant hybrids is beyond the aim of this thesis, but some of this material has been used during the analysis and will be described in the next chapters of this introduction.

During the second half of the last century, European regulation about wine limited the use of the hybrids for wine production, limits that became more stringent during the last decades. Many European

countries did not allow for long times to sell wine produced from hybrid varieties. Furthermore, the advent in the 50s of new agricultural techniques, machines (tractors), chemical pesticides and herbicides apparently provided a viable solution to the issues caused by grape pathogens; as consequence the research and the production of hybrid grape varieties dropped quickly to leave room to the cultivation of highly productive and qualitative *vinifera* varieties.

1.2 Current trends in grape cultivation and wine production

Recently, grape cultivation is suffering due to the scarce resistance of the *Vitis vinifera* varieties. The general lack of variability, due to domestication and vegetative propagation, makes *Vitis vinifera* susceptible to the pathogens and to climate change, implying higher need of treatments for the grapes to be grown.

In the last decade, the massive use of pesticides and herbicides correlated with the development of several human diseases including cancer, cardio-circulatory diseases, and neurodegenerative diseases (Baldi, 2003). The use of many pesticides related to diseases has been prohibited from the authorities, e.g. Organophosphates. Nevertheless, the scarce accountability of the farmers on the use of chemical products on cultivated areas, forces the authorities to monitor continuously the water and the fields. Furthermore, in some areas, massive use of fungicides is leading to fungicide resistance of the Powdery mildew (Gadoury et al. 2012), leading to an increase in pesticide use. Pollution is already a problem in very populated areas, and a general decrease in the use of chemicals for plant cultivation is recommended by Food and Agriculture Organization (FAO). In the 1970s the American Food and Drug Administration (FDA) developed a program called the Integrated Pest Management (IPM), that aims to reduce (if not eliminate) the use of pesticides in crop production. This program is approved and taken as a model by FAO. Among the different integrated strategies to fight pests without chemicals, the use of resistant varieties is encouraged to reduce naturally the need of pesticides (http://www.epa.gov/opp00001/factsheets/ipm.htm).

In grape cultivation, the production of novel and resistant varieties can be achieved by interspecific-crossings and/or genetic modification (Myles 2013, Borneman et al. 2013). In the last two decades the genome of *Vitis vinifera* has been sequenced (Jaillon et al. 2007, Velasco et al. 2007, Myles et al. 2010), its genotyping is straightforward (Emanuelli et al. 2013, Zarouri et al. 2015), numerous metabolic pathways have been discovered and their regulation is continuously revealed (Boss et al. 1996, Kobayashi et al. 2004, and many others). The recent advances in Genomics, Transcriptomics and Proteomics tools enhanced the level of the studies and the research on novel grape

hybrids, allowing in the near future to quickly moving forward in the development of more resistant grape varieties.

The effort in the direction of improving grape quality is a much more long-term, since this has to be ascertained in the final products; in case of fermented or refined grape products (e.g. wine, spirits), the final quality cannot be yet predicted from the starting material (i.e. there is not a complete understood route from good grape to good wine). In Wine production, several Milestones exist and they are respected by winemakers, but deeper research on the chemical grape berry composition, yeast nature and use, wine storage, aging and many other parameters is required.

1.3 From the grape metabolome project to my project

In our lab, trying to reduce the lack of knowledge about grape and its derived products, we aimed to fill the gap about the chemical composition of the grape berries, performing a huge experiment called "grape metabolome". In this experiment, the grape berries of over 100 different genotypes have been analyzed for four years with seven different analytical platforms (Mattivi et al. unpublished data). I contributed actively in both instrumental and data analysis of this experiment; but the thesis will not cover the outcome of the grape metabolome project.

In this thesis, the data obtained from the grape metabolome has been used to formulate hypotheses regarding characteristics that have not been described yet in literature. Multiple hypotheses have been generated, but I decided to focus on two of them, after considering their feasibility and the timing of the planned confirmative experiments. Both hypotheses are outcome of mere observation of the raw chromatograms of the grape samples.

First hypothesis comes from the observation that some varieties, especially the aroma-rich ones, showed a higher number of peaks with uncommon retention times and rather high relative mass defect. My hypothesis was that these peaks were aroma-precursor compounds (glyco-conjugated volatiles) and I designed an innovative approach to identify them; the experiment assessing their identification and quantification is described in chapter 5.

The second hypothesis comes from the observation that grape species from America show a rather different metabolic profile in comparison to the domesticated *Vitis vinifera*. The difference was consistently wide between *Vitis vinifera* and all the American species studied, while in literature the differences reported were multiple but not so wide. My hypothesis was that analyzing singularly the tissues of the berries of the diverse grape species, I could highlight sharper variations between the wild American grapes and the domesticated *Vitis vinifera.* This experiment is described in chapter 7.

Last, but not least, working with such wide dataset of samples and standards used in the grape metabolome experiment, I realized that there is a possibility to create a regression model able to quickly classify the signals obtained from the samples. The regression model could be used to narrow, during markers identification, the number of possible structures compatible with markers' RT, MS and MS/MS spectra. The regression model built is described in chapter 6.

1.4 Aim of the thesis

Given the gaps in the knowledge on the *Vitis* germplasm from a metabolic point of view and its importance from the quality of the grape and its derived products, this thesis aims to evaluate the hypotheses generated observing the grape metabolome data, designing and performing specific experiments to confirm or discard such hypotheses. To achieve this goal, this thesis has been divided into three major parts:

1) The development of a method for the direct identification and semi-quantification of the aroma compound precursors through LC-MS (chapter 5).
2) Establishment of a computer assisted identification method to speed up and strengthen the putative identifications (chapter 6).
3) Comparative analysis of the metabolomes of *Vitis vinifera* varieties versus some American *Vitis* species (chapter 7).

1.5 Outline of the thesis

The thesis has been organized in this way:

Chapter 2: Description of the grape materials used in this thesis

Chapter 3: Description of the analytical materials and methods used to collect the data

Chapter 4: Metabolomics: the basic concepts, experimental design and data analysis

Chapter 5: Fusion of GC/MS and LC/HRMS data to improve the identification and confirmation of the unknown Volatile-aroma-compound precursors' in Grape.

Chapter 6: Compound characteristics comparison method and its integration in the data analysis process

Chapter 7: An untargeted comparative analysis of the metabolomes of *Vitis* vinifera vs four American *Vitis* species reveals big differences in the accumulation of Polyphenols and Aroma precursors

In the abstracts of the experimental chapters (5, 6, 7) you will find a statement regarding the level of my contribution in each of the experimental part.

References chapter 1

1. Baldi, I. (2003). Neurodegenerative Diseases and Exposure to Pesticides in the Elderly. American Journal of Epidemiology, 157(5), 409–414. doi:10.1093/aje/kwf216
2. Borneman, A. R., Schmidt, S. a, & Pretorius, I. S. (2013). At the cutting-edge of grape and wine biotechnology. Trends in Genetics: TIG, 29(4), 263–71. doi:10.1016/j.tig.2012.10.014
3. Boss, P. K., K, P., Robinson, S. P., Davies, C., Robinson, S. P., Osmond, G., & Scientific, C. (1993). Analysis of the Expression of Anthocyanin Pathway Genes. *Vitis*, 111(1 996), 1059–1066.
4. Chitwood, D. H., Ranjan, A., Martinez, C. C., Headland, L. R., Thiem, T., Kumar, R., … Sinha, N. R. (2014). A modern ampelography: a genetic basis for leaf shape and venation patterning in grape. Plant Physiology, 164(1), 259–72. doi:10.1104/pp.113.229708
5. Emanuelli, F., Lorenzi, S., Grzeskowiak, L., Catalano, V., Stefanini, M., Troggio, M., … Grando, M. S. (2013). Genetic diversity and population structure assessed by SSR and SNP markers in a large germplasm collection of grape. BMC Plant Biology, 13, 39. doi:10.1186/1471-2229-13-39
6. Gadoury, D. M., Cadle-Davidson, L., Wilcox, W. F., Dry, I. B., Seem, R. C., & Milgroom, M. G. (2012). Grapevine powdery mildew (Erysiphe necator): a fascinating system for the study of the biology, ecology and epidemiology of an obligate biotroph. Molecular Plant Pathology, 13(1), 1–16. doi:10.1111/j.1364-3703.2011.00728.x
7. Galet P. (1952), Précis d'ampélographie pratique. Impr. P. Déhan
8. Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., … Wincker, P. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature, 449(7161), 463–7. doi:10.1038/nature06148
9. Kobayashi, S., Goto-Yamamoto, N., & Hirochika, H. (2004). Retrotransposon-induced mutations in grape skin color. Science (New York, N.Y.), 304, 982. doi:10.1126/science.1095011
10. Lamboy, W., & Alpha, C. (1998). Using Simple Sequence Repeats (SSRs) for DNA Fingerprinting Germplasm Accessions of Grape (*Vitis* L.) Species. American Journal of the Society of Horticultural Sciences, 123(2), 182–188.
11. Myles, S. (2013). Improving fruit and wine: what does genomics have to offer? Trends in Genetics : TIG, 29(4), 190–6. doi:10.1016/j.tig.2013.01.006
12. Myles, S., Boyko, A. R., Owens, C. L., Brown, P. J., Grassi, F., & Aradhya, M. K. (2011). Genetic structure and domestication history of the grape. doi:10.1073/pnas.1009363108/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1009363108
13. Myles, S., Chia, J. M., Hurwitz, B., Simon, C., Zhong, G. Y., Buckler, E., & Ware, D. (2010). Rapid genomic characterization of the genus *Vitis*. PLoS ONE, 5(1). doi:10.1371/journal.pone.0008219
14. Negrul AM, Baranov A, Kai YF, Lazarevski MA, , Palibin TV, Prosmoserdov NN. "Origin and classification of cultivated grape" In The Ampelography of the USSR. Eds. Moscow: 1946, 159–216

15. This, P., Jung, a, Boccacci, P., Borrego, J., Botta, R., Costantini, L., … Maul, E. (2004). Development of a standard set of microsatellite reference alleles for identification of grape cultivars. TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik, 109(7), 1448–58. doi:10.1007/s00122-004-1760-3
16. Velasco, R., Zharkikh, A., Troggio, M., Cartwright, D. a, Cestaro, A., Pruss, D., … Viola, R. (2007). A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PloS One, 2(12), e1326. doi:10.1371/journal.pone.0001326
17. Viala P. & Vermorel V. (1910), Ampelographie, traité général de viticulture. Editions Jeanne Lafitte
18. Zarouri, B., Vargas, A. M., Gaforio, L., Aller, M., de Andrés, M. T., & Cabezas, J. A. (2015). Whole-genome genotyping of grape using a panel of microsatellite multiplex PCRs. Tree Genetics & Genomes, 11. doi:10.1007/s11295-015-0843-4

2. Grape materials

2.1 *Vitis vinifera* grapes

Seven different *Vitis vinifera* varieties were chosen for this thesis. All the grapes have been collected at 18° ± 0.5 brix, and immediately frozen under liquid Nitrogen. The *vinifera* here selected can be divided in three groups according to their diffusion in the market.

Group 1: Internationally recognized varieties: Merlot, which is probably the most cultivated red grape in the world, Sauvignon Blanc and Riesling, that are respectively the second and the third most cultivated white varieties in the world (after Chardonnay).

Group 2: Local famous varieties: Gewürztraminer is a mutation of the original Traminer from Tramin (South Tyrol, Italy), it is very famous in the market and is becoming every day more international. Muscat Ottonel is a muscat variety from southern Germany, it is cultivated only in Northern Italy, Austria and Southern Germany. Moscato Rosa variety is cultivated only in Dalmatia, Istria and North-East Italy and produces a precious wine.

Group 3: Intraspecific hybrid: Iasma Eco 3, a variety patented last year from my institution (FEM).

Here follows a brief description of the materials used in this thesis.

2.1.1 Iasma ECO 3 (ECO)

Iasma Eco 3 is a proprietary variety patented by FEM in 2014. It is an intraspecific hybrid obtained by crossing "Moscato Ottonel" x "Malvasia Bianca di Candia". Numerous hybrids have been evaluated and this one together with Iasma Eco 1 and Iasma Eco 2 has been patented. Due to its aromatic progenitors, the main characteristic of this variety is its aroma. It is very rich in free terpenols and terpendiols, as reported by Ghaste et al. (2015). Nevertheless, it has been appreciated for its high tannic content.

ECO is a very productive variety; it has a robust trunk and it is leafy. The leaves are medium size, larger than long. The berries are round of 17 x 17 millimeters, with few small seeds. The skin is thick and yellowish, while the flesh is very sweet. The main characteristic of this variety is that it is super aromatic.

Image 5: The cluster and the leaf of Iasma Eco 3. From the experimental fields of the Fondazione Edmund Mach, San Michele all'Adige, (TN), Italy.

2.1.2 Gewürztraminer/Savagnin jaune (GWT)

Gewürztraminer is a natural mutation of the original Traminer from Tramin in South Tyrol region (Italy). It is famous worldwide due to its peculiar aroma and sweetness. It is used alone or blended with other grapes to produce numerous quality aromatic wines. This variety is rich in aromatic compounds, both in free and bound forms. The wine produced from this grape is very aromatic and spicy. It is mostly cultivated in the Italian Tyrol region, in Alsace (France), in Austria, southern Germany, Hungary, Slovenia and Croatia. As you may notice from the pictures, two different forms exist: the yellow one considered as the original Traminer, and the pink one, the Gewürztraminer, but they are considered genetically almost identical. Even though, the genetic causes of the different berry colors are unknown in this variety, both yellow and pink Gewürztraminer are supposed to be the same variety, with a genetic mutation occurred in only a little part of its genome. In some Gewürztraminer clusters, yellow berries can be found, meaning that the color mutation can be reversed. In my experiments, I used the Gewürztraminer pink berries.

GWT is a medium vigorous variety with medium trunk, medium size leaves and small berries, with a skin color variable from yellow to pink, and with visible lenticels. The seeds are small and between 2 to 4 in each berry. The flesh is not very consistent, with a sweet, slightly acidic taste, and a very aromatic note.

Image 6: The grape cluster of Gewürztraminer, also known with the name Savagnin Jaune. The sketch on the left is from the book "Ampelographie" of Pierre Viala and Victor Vermorel (1910). The picture on the right is from the *Vitis* international variety catalogue (www.vivc.de).

2.1.3 Merlot (MER)

The origin of Merlot is unknown; it is believed to be from Bordeaux during the first half of the XIX century or earlier. The name is because the local blackbird (Merleu in Occitane) loves to eat this grape. Merlot has been demonstrated to be an intraspecific hybrid of "Cabernet Franc" and "Magdeleine Noire des Charentes" (Boursiquot et al. 2009). Nowadays, it is widespread in the world, and is one of the most used varieties to produce quality wines. Together with "Cabernet sauvignon", "Malbec" and "Petit Verdot" is one of the varieties used to produce "Bordeaux" wines. The variety is mostly used in two ways: new world producers tend to follow a late ripening to increase the color and the tannic taste of the variety, while Bordeaux producers (classic producers) harvest the grape earlier to exploit its acidity and aroma.

MER is a vigorous variety, with a hard trunk and hard petioles. The leaves are very big, larger than longer. The berries are in average 20 x 20 millimeters, usually spherical with a very dark skin. Need are medium size 5 x 5 millimeters and the flesh is sweet, slightly acidic, with few aromatic notes.

Image 7: The grape cluster of Merlot. The sketch on the left is from the book "Ampelographie" of Pierre Viala and Victor Vermorel (1910). The picture on the right is from the *Vitis* international variety catalogue (www.vivc.de).

2.1.4 Moscato Rosa (MOR)

The origins of Moscato Rosa are likely Greek. Across the centuries it spread in the entire Adriatic coast up to Slovenia and then arrived to Trentino and Friuli regions (Northern-East Italy), where it is, nowadays, mostly cultivated and vinificated. It is part of the Muscat family, and its second name is due to its rose aroma (Rosa = rose). Moscato Rosa variety is not very famous, but the raisin wine produced with this grape is one of the most appreciated from the amateurs. The main defect of this variety is that it is not very productive mostly due to the Millerandage, which is often found in its clusters.

The plant is of medium dimension with a robust trunk and medium size leaves. The berries are of medium size (18 x 18 millimeters), with medium size seeds in a number of four per each berry. The skin is quite dark (bluish mostly), and the flesh is very sweet and aromatic. It is a late ripening grape and it is generally dried to obtain the raisins used to produce its sweet aromatic raisin wine.

Image 8: The grape cluster of Moscato Rosa.

2.1.5 Moscato Ottonel (MOT)

Moscato Ottonel (MOT) is a hybrid of "Chasselas" x "Muscat de Saumur" obtained in the late XIX century in France. It is part of the Muscat family, and, as every Muscat, it has a strong aromatic note. It is cultivated in Alsace, Southern Germany, Romania and Serbia, and is used to produce sweet and aromatic wines.

The plant is not very big, with a medium size trunk, many leaves of medium to small leaves of 11 x 11 centimeters. The berries are round, medium size (15 x 15 millimeters); the seeds are small of 4 x 4 millimeters, in a number of 2-3 in each berry. The skin is thin, with a yellow to greenish color, with few dark lenticels. The flesh is sweet and quite aromatic, with notes of peach and musk.

Image 9: The grape cluster and two leaves of Moscato Ottonel. (pictures from the *Vitis* international variety catalogue, www.vivc.de)

2.1.6 Riesling (RIE)

Riesling is a variety originally from Germany, where nowadays it is mostly cultivated and vinificated. Its origin is unknown but it is considered one of most antique variety in the world. It is the 20th most planted variety, and the third among the white ones, after "Chardonnay" and "Sauvignon Blanc" (OIV report, 2014, www.oiv.int). It is cultivated in the central Europe, from North Germany to Northern Italy, from Eastern France to Hungary. It is also cultivated in NC and Zeeland, Australia and Chile. Wines produced with this variety have a big influence from the "terroir" of the cultivated zone. Their peculiar characteristic is its acidity mixed with citrus, apple and peach aroma.

RIE is a variety of moderate growth, with a robust trunk and many leaves. Its leaves are medium size, usually larger than longer. The berries are usually small (12 x 12 millimeters), with a yellow to greenish color. The skin is very thin, seeds are medium size (5 x 3.5 millimeters) in a number of 2-3 per berry. The taste is mostly acidic, but this variety can achieve very long maturation periods (late maturing), that enables it to accumulate a lot of sugar and specific aromas.

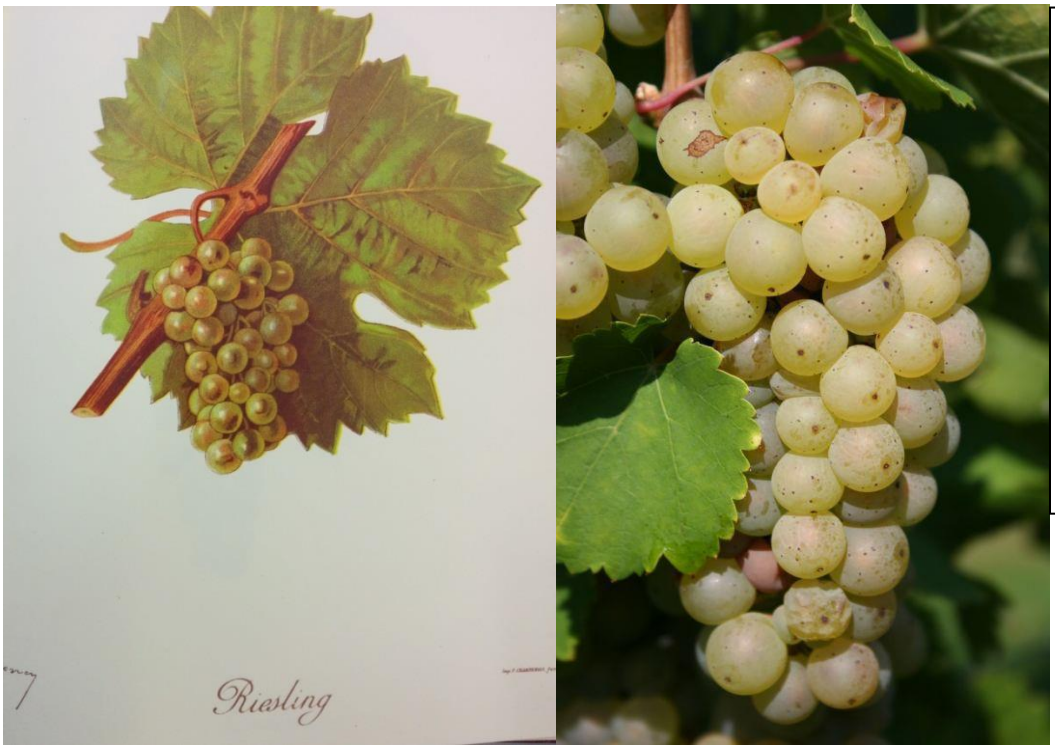


Image 10: The grape cluster and leaves of Riesling variety. The sketch on the left is from the book "Ampelographie" of Pierre Viala and Victor Vermorel (1910). The picture on the right is from the *Vitis* international variety catalogue (www.vivc.de).

2.1.7 Sauvignon Blanc/Gros Sauvignon (SAU)

The "Gros sauvignon" or "Sauvignon blanc" has been confused with the "Sauvignon" for decades; indeed settlers were thinking that the difference was due to the different areas of cultivation

and not in a varietal difference. It is a very famous variety planted in Italy and France, USA, Chile, South Africa, Australia and New Zeeland. The grapes bud late, but ripe early; they are usually quite acidic with peculiar aromas due to methoxy-pyrazines and volatile thiols. The oak aging is common for the wine produced from this grape: it helps in rounding the aromas and reduces the acidity.

The leaves of the SAU are quite big, as well the berries that reach the size of 20x20 millimeters. The seeds are rather big with a size of 6x6 millimeters. The skin is very thin, the color is yellow to greenish and it is sweet and a bit aromatic. Wines produced with this variety are not very alcoholic (11 to 13°), with an acidic taste and a slightly aromatic note.

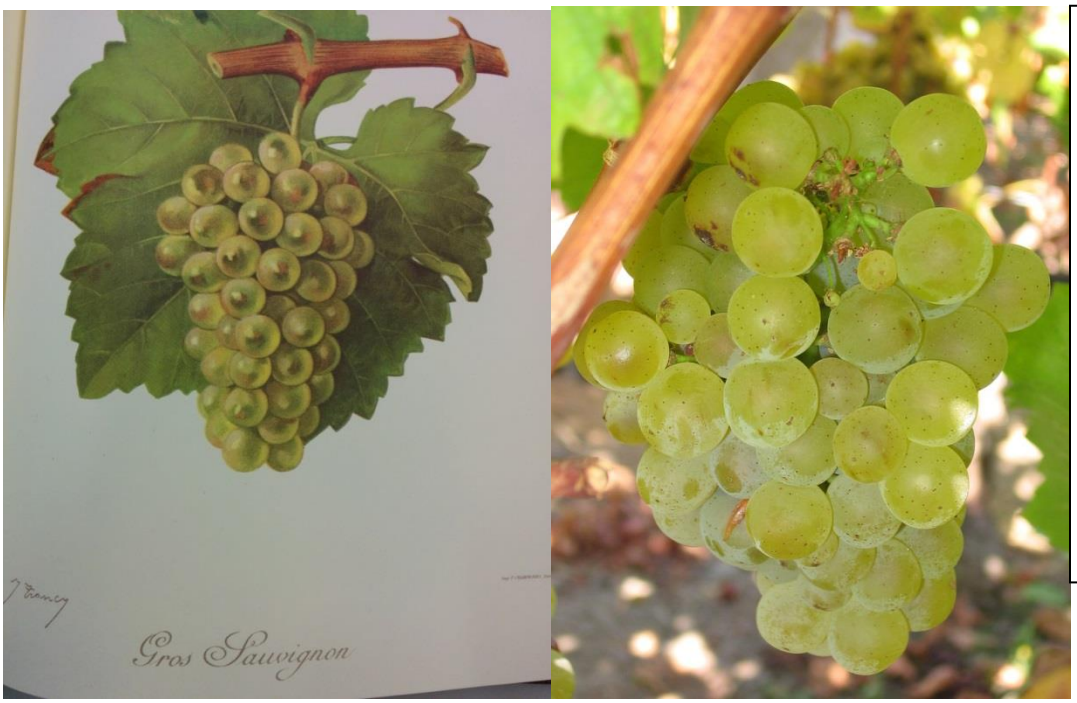


Image 11: The grape cluster and leaves of Sauvignon Blanc variety also known with the name of Gros Sauvignon. The sketch on the left is from the book "Ampelographie" of Pierre Viala and Victor Vermorel (1910). The picture on the right is from the *Vitis* international variety catalogue (www.vivc.de).

2.2 The American *Vitis* germplasm: general considerations

American *Vitis* germplasm encloses many different eco-species, mostly original from Southern and Eastern United States and Mexico, where they still grow as wild plants. Few species growing in Canada exist, like *Vitis aestivalis* and *Vitis riparia* that are now widespread in the entire eastern coast, from Ontario (Canada) down to Florida (USA), indicating their resistance to very different climates.

The main agricultural importance of the American *Vitis* is due to their resistance to many pests; for example, as stated above, the main solution to the XIX century "Phylloxera" blight was the use of resistant American rootstocks to graft susceptible *Vitis vinifera* varieties. Indeed American *Vitis* are resistant to many pests that nowadays are infesting the *vinifera* vineyards, especially Powdery mildew, etc.

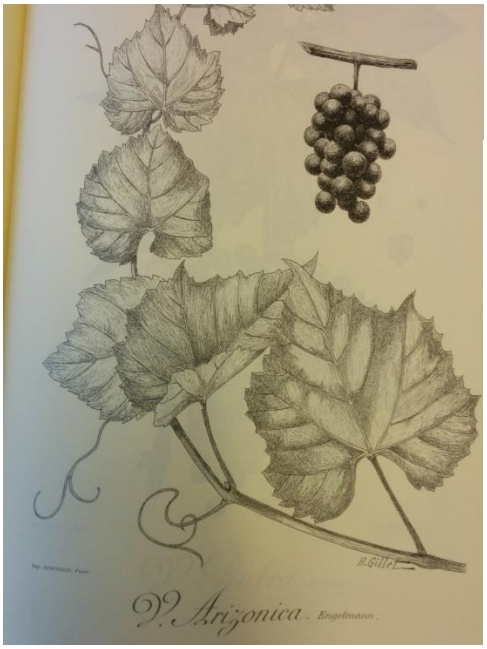
American *Vitis* germplasm is very vast; the natural barriers between the different regions of the northern America speeded up the speciation process, while since the XVI century they have been used for cultivation and hybridization by American settlers (*Vitis vinifera* was not able to grow in northern America). Many wild species exist, and many varieties and hybrid varieties has been established during the last four centuries. An impulse to hybrid establishment has been the spread of diseases, especially in the XIX century in France, where breeders established numerous interspecific hybrids, commonly called French-American hybrids. If the first hybrids established had poor quality in comparison to *Vitis vinifera* varieties, nowadays many valuable crosses are cultivated, especially in the USA (http://www.hort.cornell.edu/reisch/grapegenetics/cultivars.html). Some of the hybrids have no *vinifera* parentage, but still are able to produce quality grapes.

Here, I will describe only the material used in my experiment, taking into account that it is only a very little part of all the variability existing within the American *Vitis* germplasm. Our institute is rich in wild *Vitis* species, nevertheless the choice of the American grape materials was limited to the ones analyzed in the grape metabolome project (Mattivi et al. unpublished data). I chose only the material that was able to produce grape berries in a sufficient amount with the desired characteristics (18° brix). The classification and the pictures displayed in the next chapters are from the book series "Ampelographie" published by Pierre Viala and Victor Vermorel (1910).

2.2.1 *Vitis arizonica Texas*

Vitis arizonica has been classified by Engelmann in 1868; in nature, it is original to the whole Arizona, some mountain regions of Northern Mexico, Nevada and some western Texas. It is very robust species, adaptable to very big temperature shift. It has been found growing in places up to 8000 feet altitude (2700 meters). *Vitis arizonica Texas* is a phenotype coming from western Texas. Generally, this species is very resistant to Chlorosis, but it is not very resistant to Phylloxera and is very rich in nodes. For these reasons, it cannot be used as rootstock for *Vitis vinifera* plantations (Ravaz, 1902)

The plant has strong roots and a strong thin ligneous trunk, while the leaves are small, with small spherical berries of diameter and longitude of 7 x 7 millimeters. Seeds have a diameter and longitude of 3.5 x 4.5 millimeters. The skin is dark and thick, flesh is very soft with a prominent acidic taste.



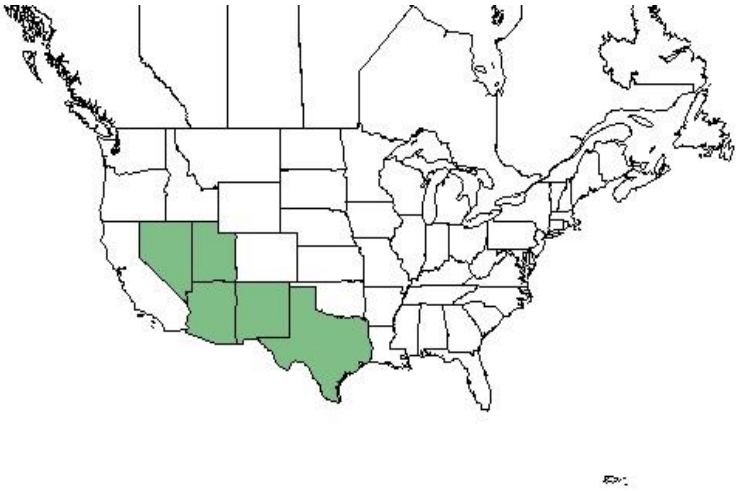
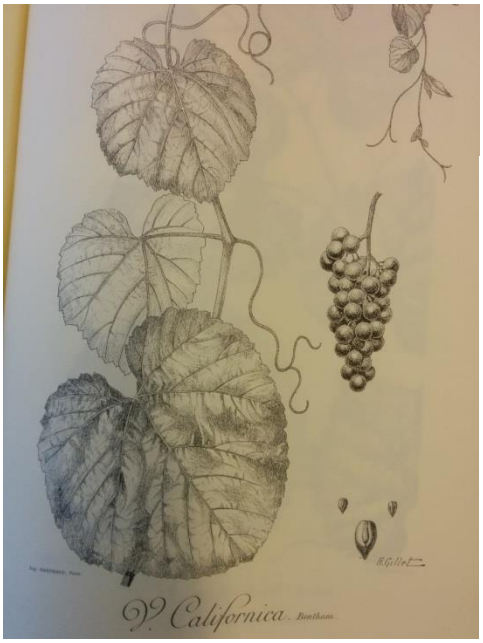


Image 12: On the left, the grape cluster and leaves of the *Vitis* arizonica. The sketch is from the book "Ampelographie" of Pierre Viala and Victor Vermorel (1910). On the right, the areal of diffusion of the *Vitis* arizonica that goes from Nevada to Texas. (data from the National resources conservation services, http://plants.usda.gov/core)

2.2.2 *Vitis* californica

Vitis californica has been classified by Bentham in 1844. It is original from California and Oregon states. It is mostly found close to the rivers of the slopes of the mountain range to the pacific coast. It has vigorous roots, with a wide trunk and many long secondary ramifications. It is not very resistant to the Phylloxera and it suffers the Powdery mildew more than *Vitis vinifera*. For these reasons, it was of scarce interest for breeders, and it has not often used to constituted hybrids with *Vitis vinifera* (Ravaz, 1902)

The leaves are quite big (22 x 22 centimeters) and produces small berries with both diameter and length of 7 millimeters, and seeds of 3 x 6 millimeters of diameter and length. The skin is very dark, flesh is consistent with a sweet taste.



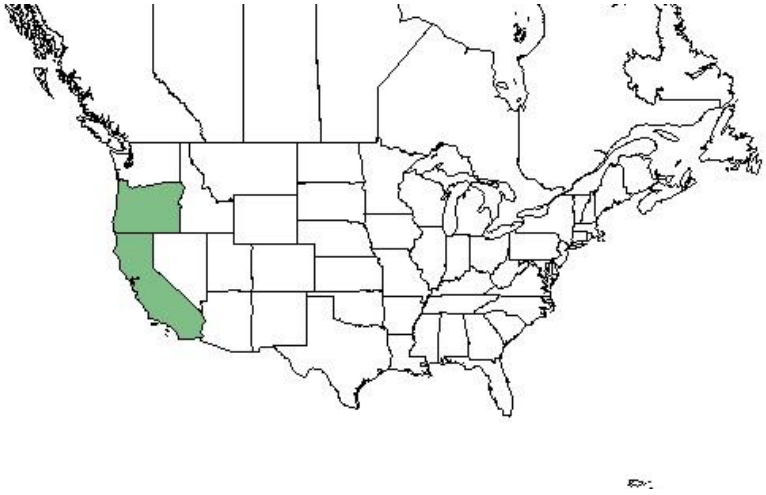


Image 13: On the left, the grape cluster and leaves of the *Vitis* californica. The sketch is from the book "Ampelographie" of Pierre Viala and Victor Vermorel (1910). On the right, the areal of diffusion of the *Vitis* californica that goes from Oregon to California. (data from the National resources conservation services, http://plants.usda.gov/core)

2.2.3 *Vitis* cinerea

Vitis cinerea has been first described by Engelmann in 1867, even if he classified it as sub-species of *Vitis aestivalis.* Was only Millardet that in 1878 proposed *Vitis cinerea* as a different species. The main confusion about this species is that it is widespread in the central states of United States, from Georgia to Texas, from Missouri to northern Mexico, and it is difficult to find wild forms, while hybrid with other *Vitis* species are very often found in nature especially in Missouri valley, where it cohabits with *Vitis cordifolia. Vitis cinerea* is very resistant to Phylloxera, and to the fungal pathogens. It is susceptible to chlorosis but it grows very well in Silicon-rich fields (Ravaz, 1902)

It has a very strong trunk, of about 40 centimeters in diameter, and very long branches. It can have leaves very different plant by plant (from 7 to 20 centimeters in both diameter and length). The berries are small of spherical form (7 x 7 millimeters) with seeds long 5 millimeters and large 2.5, often unique in the berry. The skin is lucent and dark, the flesh is greenish with an acidic taste.

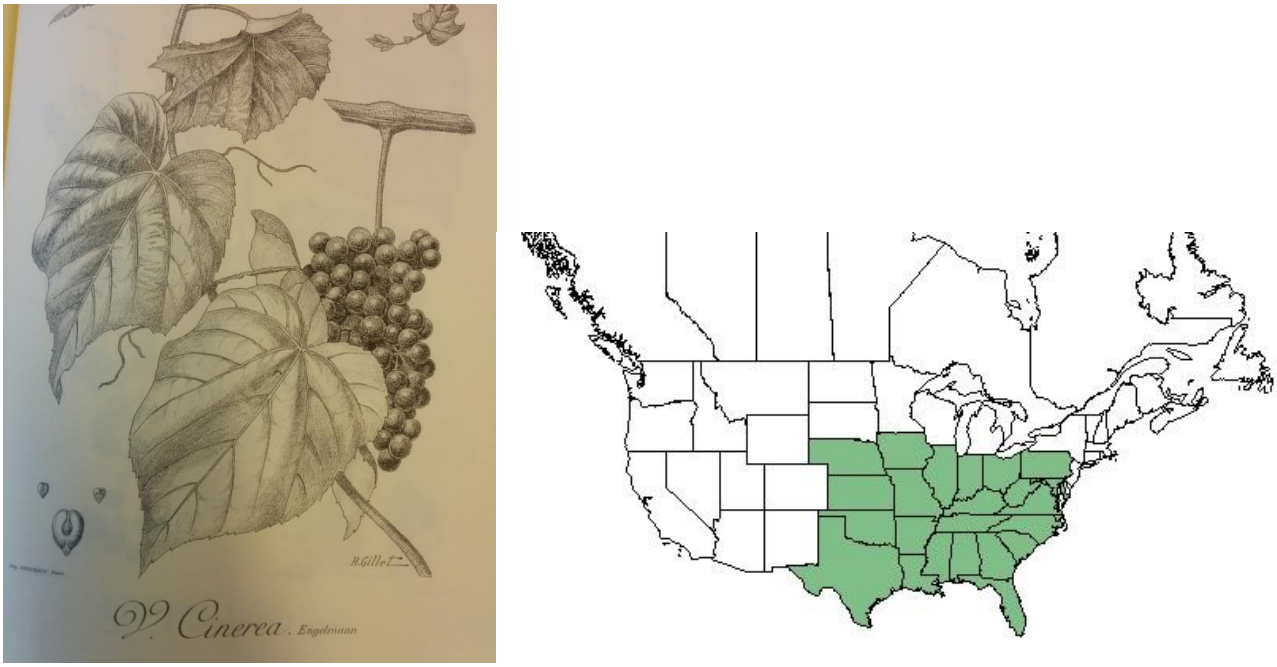


Image 14: On the left, the grape cluster and leaves of the *Vitis* cinerea. The sketch is from the book "Ampelographie" of Pierre Viala and Victor Vermorel (1910). On the right, the areal of diffusion of the *Vitis* cinerea that goes from central to eastern United States. (data from the National resources conservation services, http://plants.usda.gov/core)

2.2.4 Vitis berlandieri x Vitis riparia Teleki selection "Kober 5 BB" (K5BB)

K5BB is a hybrid variety obtained around 1910 in Austria by crossing *Vitis berlandieri* with *Vitis riparia.* It is commonly used as rootstock because it showed to have a good resistance to Phylloxera and to active lime in the soil. It does not suffer temperature changes and lack of calcium and potassium in the soil. It might have inconstant production. Both of its parents are original from USA: *Vitis berlandieri* was firstly found in the northern area of Texas, while *Vitis riparia* is from Missouri and Mississippi. Hybrids of this two species can be found in nature, being inter-fertile and widespread in the same regions (Ravaz, 1902).

As selected rootstock, K5BB has a very robust trunk, with strong roots. Leaves are medium size (15 x 15 centimeters), with the berries that are small (7 x 7 millimeters) and seeds of 4 x 5 millimeters of diameter and length. Its skin is very thick and the flesh is dark red colored with an acidic taste.

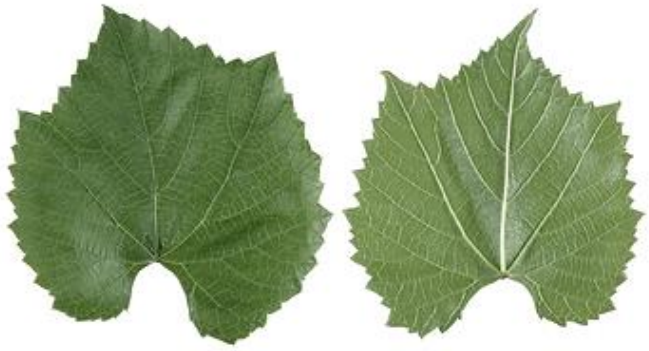


Image 15: The leaves of the K5BB. The picture is from the UC Davis grape varietal collection database:

http://iv.ucdavis.edu/Viticultural_Information/

2.3 Hybrid varieties

As stated earlier, the major part of the *Vitis* genus is composed of inter-fertile species; crosses between the different species happen in nature, and are very common in Northern America, where multiple wild species co-exist. Many breeders took advantage of this characteristic to create hybrid varieties crossing different species. From a commercial point of view, hybrids of wild *Vitis* have poor or none economic value and will not be considered in this work. In this thesis, for "Hybrid varieties" I refer to all the varieties obtained by crossing *Vitis vinifera* with another *Vitis* species.

The hybrids obtained from selected *Vitis vinifera* parents give a higher value and many interesting hybrids have been obtained during the last century. On the other hand, the non-*vinifera* heritage often gives the desired robustness and resistance to the hybrid progeny. Sounds obvious that a commercial variety is not obtained after only one-step hybridization, and that further steps of backcrossing or crossing with other hybrids/species are needed to obtain a variety that has the desired characteristics. Nevertheless, the description of the whole hybrids set is out of the aim of this thesis that wants only to focus on the common hybrid selection procedure, to point out the limits of the hybrids and to describe the materials used in this work.

The first hybrids have been established by American settlers, but it was only in the second half of the XIX century, during the "Phylloxera" grape blight, that hybridization became a common practice. In Europe, susceptible *Vitis vinifera* were crossed with American *Vitis* to obtain hybrids resistant to Phylloxera but also resistant to lime. Indeed American *Vitis* were not accustoming to lime-rich soil that in Europe are very common. Hundreds of hybrids have been produced since 1870, and up to 1950, they were commonly planted, grown and used for vinification (Pee-laby, 1929), especially in France (indeed these hybrids are commonly called French-American hybrids).

Due to the introduction of new agricultural practices in the late 40s and 50s, like new mechanical instrument, new pesticides and herbicides, the interest in resistant hybrids decreased. Furthermore, around 1940 Europe started to prohibit the commercialization of wines produced from hybrid grapes. Only *Vitis vinifera* grapes have been allowed to produce wine for decades.

Despite of the prohibition, breeders mostly from Germany, Switzerland, Austria, Hungary and Czech Republic continued to create hybrids starting from the old French-American hybrids, with the aim to confer resistance, good yields and a quality comparable to *Vitis vinifera* varieties (Bouquet et al. 2000). Nowadays, hundreds of hybrids have been established, most of which are resistant to the main grape pests. A specific mention is given to the PIWI grapes (Pilzwiderstandsfähige = fungus resistant),

that as the German word states are grapes resistant to fungus infections. Many of these hybrid varieties exist and are now commercialized both as wine and table grapes.

As stated in the previous chapter, the cultivation of susceptible *Vitis vinifera* varieties is no longer sustainable, due to the high production and environmental cost of their plantation. Many fruit plant breeders are creating hybrid resistant varieties that are cultivated and sold in fruit market. Therefore, to reduce costs and improve the environmental sustainability, hybrids cultivation will be a common practice in the next years. From this point of view, the study of the germplasm resources is basic to understand the possibilities that hybrid varieties give to breeders and to shorten the hybridization process, selecting the adequate material and using marker assisted selection to select the promising progenies in hybridization programs.

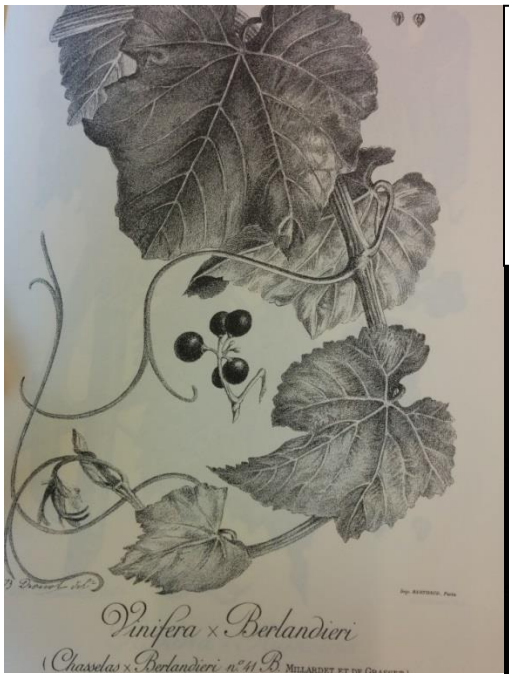
In the next paragraphs, I will describe the three hybrid varieties chosen in this experiment. The first two are interspecific crossing with 50% *vinifera* heritage and 50% American heritage. The third choice is Nero, one of the first PIWI grape established back in the 1965. Their inclusion in the experiment was not to compare their characteristics to the *Vitis vinifera* varieties (otherwise, more recent hybrids would have been chosen). The idea was to see if these first stage hybrids (F_1) tend to resemble more one of the two groups from a metabolic point of view. This was intended to see which desirable and undesirable metabolic characteristics tend to be transmitted to the progeny. On the other hand, Nero was chosen to see if a patented variety already used for vinification and selected without the use of any molecular/metabolic marker possess such desirable or undesirable characteristics.

2.3.1 Chasselas x *Vitis* berlandieri 41B (Millardet & De Grasset)

41B is a hybrid variety obtained by crossing the variety *Vitis vinifera Chasselas* with the *Vitis berlandieri* clones imported in France in the 1880 by Millardet. It can be considered one of the first French-American hybrids. The idea of these crosses was to obtain hybrid varieties resistant to "Phylloxera" and to the Chlorosis. 41B was obtained in 1882 and it started to be used as rootstock in 1894 to rebuild the vine plantations in the "champagne de Cognac". It is very resistant to Phylloxera, chlorosis, grows very well in cold regions (like Northern France), and has a strong trunk. Like *berlandieri* grape, it delays the fructification of the varieties grafted on. Nevertheless, it is still considered a good rootstock, especially in cold areas.

It has a robust trunk, superficial roots, leaves wider than longer, the berries are rather small (10 x 10 millimeters) with seeds similar to the ones of the *Vitis berlandieri* of size 3.5 x 4.5 millimeters. The

skin is hard, flesh is pale green and the taste is sweet. Due to its Chasselas progenitor it is also slightly aromatic.



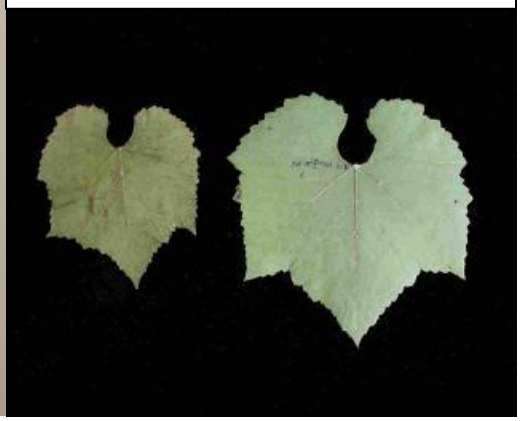


Image 16: On the left the berries and the leaves of 41B. The sketch is from the book "Ampelographie" of Pierre Viala and Victor Vermorel (1910). On the bottom, the leaf of 41B, picture from The *Vitis* internation Variety Catalogue, www.vivc.de.

2.3.2 Isabella (*Vitis vinifera* x *Vitis Labrusca)*

The origin of Isabella grape is unknown. It was believed to be a *Vitis labrusca* variety for many decades, but its different characteristics like the extreme fertility, production, susceptibility to powdery mildew and black root indicate a *Vitis vinifera* heritage in its genome. Its name is derived from Miss Isabelle Gibbs that first disclosed such seeds to the settlers. DNA analysis through molecular markers demonstrated it to be a hybrid between *Vitis vinifera* and *Vitis labrusca* (Emanuelli et al. 2013), probably obtained by chance while American cultivators were trying to cultivate some *Vitis vinifera* varieties in the United States. It is probably from South Carolina, but due to its extreme fertility, it spread quickly across the USA. It is the first "American" variety imported in Europe, and it is believed that through some of these imports "Phylloxera" arrived to Europe.

It is quite resistant to Phylloxera and grows very well in warm regions and sub-tropical areas (southern Italy, northern Africa, Corea, and China). It was exported to Europe, Africa and Asia, and nowadays it is cultivated in many countries in the world. Its foxy taste is a heritage of the *Vitis labrusca*, and it is not considered a variety for quality wines. On the other hand, its strawberry aroma is the main characteristic known to the consumers and made it famous worldwide. For this reason it is

also called “Strawberry Grape”, “Framboisier” (French), and “Uva Fragola” (Italian). It has big leaves with only three lobes; the berries are globose of with a diameter of 20 millimeters and length of 25. In each berry, there are from one to four seeds that are rather small. The flesh is consistent and has the typical *labrusca* texture, with a sweet taste and strawberry aroma due to the high content of furaneol in its flesh (Zabetakis et al. 1999).

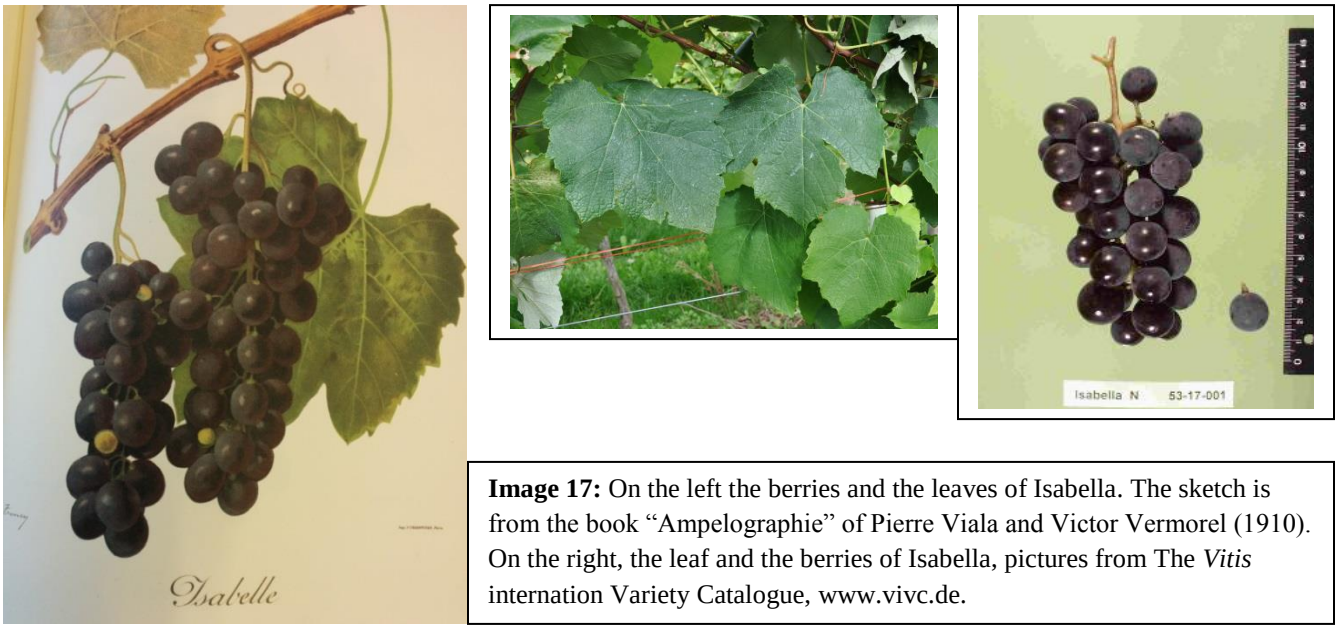


Image 17: On the left the berries and the leaves of Isabella. The sketch is from the book “Ampelographie” of Pierre Viala and Victor Vermorel (1910). On the right, the leaf and the berries of Isabella, pictures from The *Vitis* internation Variety Catalogue, www.vivc.de.

2.3.3 Nero

Nero is a rather recent hybrid varieties obtained after numerous backcrossing with *Vitis vinifera* varieties. It has in its pedigree both *Vitis berlandieri* and *Vitis rupestris* but most of its progenitors are *vinifera* grapes. Despite of this *vinifera* heritage, it still has many characteristics of its wild progenitors, like di-glycosidic bond anthocyanins, and resistance to many pathogens. It has been obtained by two Hungarian breeders “Jozsef Csizmazia” and “Laszlo Bereznai” back in 1965, inside an experimental breeding program based on the use of old French American hybrids (Csizmazia & Bereznai 1968). It was registered as wine grape only in 1993.

Nero is a resistant variety, it is a Hungarian PIWI grape (fungus resistant) and does not need the use of pesticides for its cultivation. It has wide leaves and produces spherical berries of 25 millimeters diameter. The seeds are rather big with a diameter of 7 x 7 millimeters. The skin is dark red and the flesh is greenish, but can have a red external layer if damaged by cold. The taste is sweet, and rich in tannins. It is used to produce sweet red sparkling wines.

Image 18: The shooting buds of the Nero leaves. The picture is from the *Vitis* international variety Catalogue www.vivc.de.

References chapter 2

1. Bouquet, A.; Pauquet, J.; Adam-Blondon, A. F.; Torregrosa, L.; Merdinoglu, D.; Wiedemann-Merdinoglu, S. (2000). Towards obtaining grapevine varieties resistant to powdery and downy mildews by conventional breeding and biotechnology. Bulletin de l'OIV 2000 Vol. 73 No. 833-834 pp. 445-452
2. Boursiquot, J.-M., Lacombe, T., Laucou, V., Julliard, S., Perrin, F.-X., Lanier, N., … This, P. (2009). Parentage of Merlot and related winegrape cultivars of southwestern France: discovery of the missing link. Australian Journal of Grape and Wine Research, 15(2), 144–155. doi:10.1111/j.1755-0238.2008.00041.x
3. Csizmazia J., Bereznai L. (1968): A szılı Plasmopara viticola és a Viteus vitifolii elleni rezisztencia nemesítés eredményei. Orsz. Szıl. Bor. Kut. Int. Évkönyve, Budapest. 191-200.
4. Emanuelli, F., Lorenzi, S., Grzeskowiak, L., Catalano, V., Stefanini, M., Troggio, M., … Grando, M. S. (2013). Genetic diversity and population structure assessed by SSR and SNP markers in a large germplasm collection of grape. BMC Plant Biology, 13, 39. doi:10.1186/1471-2229-13-39
5. Pee-laby E., (1929). La vigne nouvelle: les Hybrides producteurs. Librairie J.B. Baillerie et fils, Paris.
6. Ravaz L. (1902). Porte-Greffes et Producteurs-directs. Caracteres – Aptitudes. Masson et C, Editeurs.
7. Viala P. & Vermorel V. (1910), Ampelographie, traité général de viticulture. Editions Jeanne Lafitte
8. Zabetakis, I., Gramshaw, J. ., & Robinson, D. . (1999). 2,5-Dimethyl-4-hydroxy-2H-furan-3-one and its derivatives: analysis, synthesis and biosynthesis—a review. Food Chemistry, 65, 139–151. doi:10.1016/S0308-8146(98)00203-9

3. Liquid chromatography coupled to Mass Spectrometry: types, strategies and role in the separation and identification of the metabolites

Many different types of compounds co-exist in a mixture and they need to be separated to permit their identifications. The separation step is necessary for the single compound identification and the type of separation technique is chosen based on the nature of the studied compound/s. Compounds with a very different chemical nature coexist in a mixture, and their separation, at the state of the art, cannot be achieved using just one separation technique. This means that, only few classes of compounds can be separated by a unique separation technique, and to cover the whole classes set, many different approaches are required.

In untargeted studies, mass spectrometry (MS) and nuclear magnetic resonance (NMR) are the state of the art methodologies to identify and quantify large number of compounds (Dettmer et al. 2007). Their use can be coupled to chromatography, which is a separation method delaying in time the analysis of different compounds, allowing the MS/NMR to measure large number of compounds. Few chromatographic methods are available with NMR due to long analysis time required by the instrument for each compound. Furthermore, the NMR is not very sensitive, so its use in untargeted studies is limited. NMR has not been used in this thesis and will not be described here. In my thesis, I used Mass spectrometry as benchmark, and this chapter will describe the separation methodologies coupled with mass spectrometry and the basic principles behind a mass spectrometer.

3.1 Separation techniques

Mainly three separation techniques are coupled with mass spectrometry: Gas Chromatography (GC), Liquid Chromatography (LC) and Capillary Electrophoresis (CE). Moreover, some researchers prefer to have a separation of the metabolites before the analysis (off-line), extracting the interesting compounds with dedicated solvents and using then the mass spectrometer with direct infusion (DI) of the extracted mixture. A detailed description of all the four techniques is too vast to be treated in this text, and their description and comparison is delegated to the literature. A description of the basic principles of liquid chromatography will follow in the next chapter. A wider description of all the liquid chromatography techniques available nowadays can be found at http://www.americanlaboratory.com/163469-Review-of-HPLC-2014-Advances-in-HPLC-and-So-Much-More/.

3.1.1 Liquid Chromatography

Liquid chromatography (LC) is a separation technique where a liquid solvent containing the sample mixture passes through an absorbent column that attracts or repulse the mixture compounds according to their chemical nature, determining a temporal shift in their release from the column (the temporal shift is called "retention time", RT). If high pressure is used to push the compounds through the chromatographic column, the method is called HPLC (high-pressure liquid chromatography) or UHPLC (ultra-high-pressure liquid chromatography). The adequate pressure is created by pumps that pump the solvents (eluents) with a continuous flow in the instrumental tubing; an injector inserts the sample mixture in the flow, which carries the mixture to the stationary phase (column) that interacts with the eluents and sample mixture and determines the RT of the compounds. The column can be of different lengths, while the absorbent material inside can be of different materials and different particle size. The absorbent material determines the type of interaction with the compounds, while the particle size and the column size determine the number of those interactions. Stronger interactions mean longer retention times, so compounds that have a higher affinity with the absorbent material should have longer retention times. A higher amount of column/analytes interactions increases the separation of the compounds with similar (but not identical) affinity to absorbent material. Therefore, longer column and smaller particle size determine a better-separated and more resolved chromatogram.

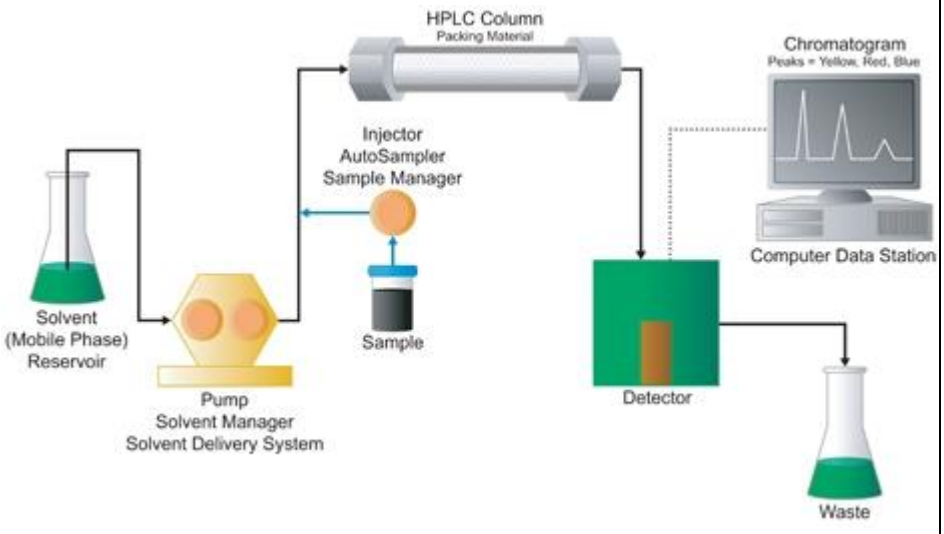


Image 1: A schematic representation of an HPLC system. A solvent (eluent) is pumped by a pump through the system. The sample is injected in the solvent flow by an injector connected to an auto-sampler. The flow is pushed through the Column, where the analytes are separated. At the end, a detector register the specific signal given by the analyte. All the signals are recorded by the computer as chromatographic peaks. Source: www.waters.com "how does liquid chromatography work?"

Numerous types of different columns exist, and depending on the aim of the analysis, the more appropriate column must be chosen. In this thesis, I only focused on the study of polar to mid-polar metabolites of the grapes; as previous researchers in our lab established a chromatographic method to

separate many mid-polar metabolites with a reverse phase column, I used such method. The column and the chromatographic method used will be described in section 3.1.1.2. In the next section, the extraction method necessary before injection will be described and discussed.

3.1.1.1 The extraction method

Many extraction methods exist to extract polar and mid-polar metabolites. Usually water in a mixture with an organic solvent like methanol, ethanol, acetonitrile or ethyl acetate is used in metabolomics, because they can be injected directly in the LC-MS instrument (Vuckovic, 2012). In this work, the extraction of grape polar and mid-polar metabolites was performed following the slightly modified protocol established by Theodoridis et al. (2012) . Grape berries were pulverized (1 gram per sample) under liquid nitrogen; 1.2 ml of methanol, 0.8 ml of chloroform, 20 µl of internal standards (indole 3-propionic acid 80 mM, 4-stilbenol 40 mM and gentisic acid 20 mM) and 2 µl of formic acid were added to the grape powder. The extract was vortexed for 30 seconds, sonicated for 10 minutes and agitated in an orbital shaker for 15 minutes. Then it was centrifuged for 5 minutes at 5000 rpm and 5° centigrade. The upper water/methanolic phases was collected, filtered through a Millipore filter (WHATMAN 0.22 um) and injected into LC-MS instrument.

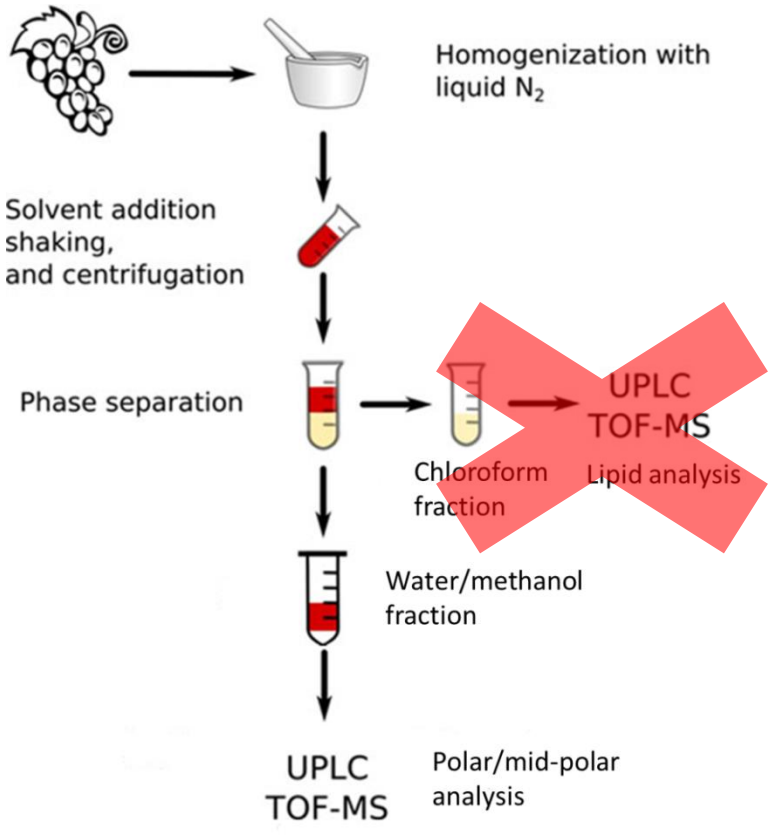


Image 2: A scheme of the extraction method used in my thesis. The grape has been ground under liquid nitrogen, extracted with Methanol/Water/Chloroform 2:1:2, 0.1% Formic acid. After vortex, sonication, and centrifugation the phases separated in two layers. The upper layer (Methanol/Water) was collected, filtered and injected in the UPLC-TOF-MS.

The method can also be used forlipids analysis, but has not been performed in my work. (Theodoridis et al. 2012)

3.1.1.2 The UHPLC instrument and the chromatographic method

A reversed phase chromatographic column consists of a non-polar stationary phase, usually consisting of straight chains of alkyl groups often bound to silica particles. The mobile phases used in this kind of columns are often polar (like water) and mid-polar (like methanol or acetonitrile); switching the gredient from to mid-polar solvents increase the solubility of mid-polar analytes trapped in the column, eluting it in different times (retention times).

The chromatographic separation of the analytes during my work on this thesis was performed using an ACQUITY UPLC system (WATERS, Manchester, UK). The instrument schematic is displayed in image 3. From the bottom to the top we can observe: A) pumps; B) the auto sampler; C) column heater/cooler; D) UV-vis PDA detector (Photo-Diode-Array). At the bottom of the instrument there are two pumps, each one dedicated to each eluent (eluent A and B); the pumping system is dedicated to the creation of the eluents flow. The second shelf is dedicated to the auto sampler, in which the samples are stored at 5° C. The samples are automatically injected in the eluent flow. The column heater/cooler is an instrument dedicate to the setting of the appropriate temperature for the column The PDA detector is a non-destructive detector that measures the UV-Vis absorbance of the analytes.

Arapitsas et al. (2014) developed the chromatographic method used in the analyses described in chapters 5 and 7. The method uses a HSS-T3 1.8 um x 2.1 x 15 cm column coupled with a dedicated pre-column. This is a silica bond column; it has an interface slightly polar that elongates the retention times of polar metabolites. This column has been used because it has proven to have a good retentivity for many different metabolites and is very stable after thousands of injections (>4000 injections). The mobile phases used in this work were Eluent A = acidified water with 0.1% formic acid and Eluent B = acidified methanol, with 0.1% of formic acid. The gradient used in the analysis was: from 0 to 1.5 minutes 100% eluent A, 10% eluent B up to 3 minutes, then an gradient up to 40% of B in 18 minutes, then up to 100% of B in 21 minutes, hold at 100% of B up to 25 minutes and then back to 100% A for a total run of 28 minutes. The column was kept at 40 °C during the analysis.

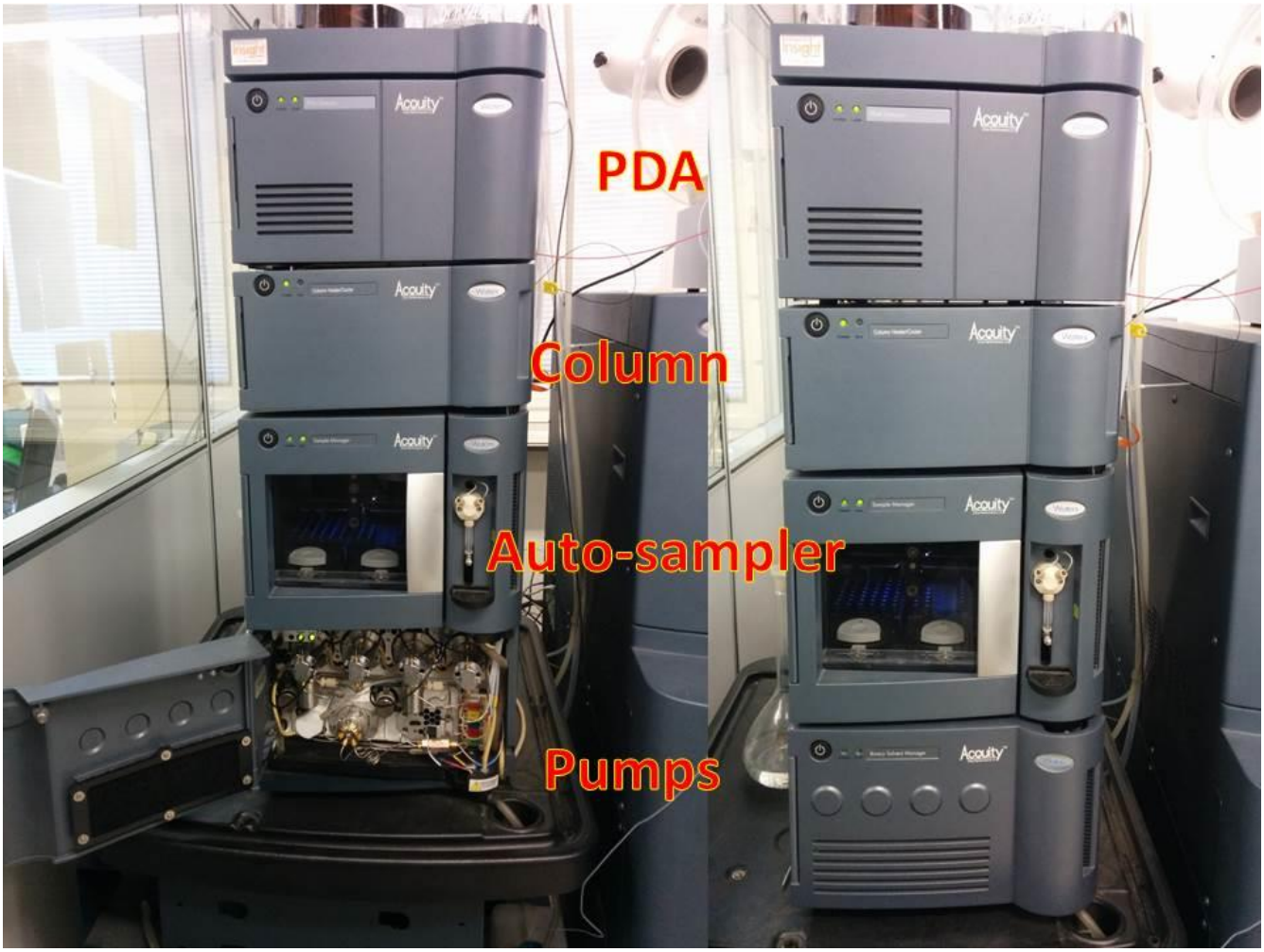


Image 3: Two pictures of the same instrument. In the picture on the left, the UPLC pumps are visible. The LC instrument is composed of three parts: the pumps, the auto-sampler and the column heater/cooler. The forth component is the UV-vis detector (Photo-Diode-Array, PDA).

3.1.1.3 The retention time: a tool to separate metabolites based on their physico-chemical properties.

The retention time of a compound is determined by the nature, strength and number of interactions that such compound is able to have with the mobile phase (eluents) and the stationary phase (column). The developed chromatographic method needs to assure the ruggedness, robustness of the analysis and the repeatability injection after injection. Whether these parameters are assured, the RT is stable and repeatable. This is a basic parameter to identify the compounds and make comparisons between samples.

During the development of a new method, we need to take in consideration that only few eluents can be coupled with ESI ion source: Water, Methanol, Ethanol, Isopropanol, Acetonitrile (I will talk

about ESI in the next section). Furthermore, numerous stationary phases are available on the market, but only few of them can be used in metabolomics analysis. In facts, the stationary phase need to assure the broadest versatility possible to analyze multiple different species of metabolites in a single run. The stationary phases used in metabolomics need to be universal for the kind of metabolites under analysis, and stable during hundreds of injections. In the metabolomics studies of polar compounds, like sugars, organic acids and amines, usually HILIC columns are used (Hydrophilic Interaction Liquid Chromatography). Indeed, Hilic technology is based on hydrophilic interactions between the analytes and the water layer in the stationary phase, increasing the separation of the analytes according to the water-affinity. Amide columns (Gika et al. 2012) and Zic Hilic columns (Zwitterionic, having both positive and negative interaction sites) are the main Hilic columns used in metabolomics. In mid-polar metabolites studies (like aromatic compounds), reversed phase columns are preferred, e.g. classical C18/silica bond columns, which assure the best separation for this kind of metabolites, with gredient moving from water to organic phases (acetonitrile or methanol.. In lipidomics studies, C18 or C30 columns are used coupled with mid-polar solvents like acetonitrile and isopropanol.

To understand how the separation of the different analytes is achieved and what the specific influence of the mobile phase and the stationary phase are, some parameters are here introduced. The main parameter determining the effect of the eluents on the analytes is their logP. LogP is the coefficient of partition of a compound in Octanol/Water mixture. It is measured as the logarithm of the amount of solute dissolved in Octanol divided by the amount of solute dissolved in water

$$\log\ P_{\mathrm{oct/wat}} = \log\left(\frac{[\mathrm{solute}]^{\mathrm{un-ionized}}_{\mathrm{octanol}}}{[\mathrm{solute}]^{\mathrm{un-ionized}}_{\mathrm{water}}}\right)$$

Across the pH range, most of the metabolites tend to assume ionized or un-ionized forms, according to their pKa. To consider this fact, the parameter logD is introduced (coefficient of distribution), that takes in account also the ionized form of the solvent/solutes. Indeed pH has a big effect on the solubility of the compounds. Therefore, the log D is measured as

$$\log\ D_{\mathrm{oct/wat}} = \log\left(\frac{[\mathrm{solute}]^{\mathrm{ionized}}_{\mathrm{octanol}} + [\mathrm{solute}]^{\mathrm{un-ionized}}_{\mathrm{octanol}}}{[\mathrm{solute}]^{\mathrm{ionized}}_{\mathrm{water}} + [\mathrm{solute}]^{\mathrm{un-ionized}}_{\mathrm{water}}}\right)$$

This equation can also be applied to the relationship between the stationary phase and the mobile phase. To give an example, if a non-polar column is used as stationary phase and polar solvents are used as mobile phases, the partition of the analytes between the two phases can be described as the logP = solute in the stationary phase / solute in the mobile phase. The ratio between the partition is obviously dependent on the ionization of the solute (logD), so changing the pH and/or the polarity of the mobile phase, quickly changes the ratio, eluting the analytes.

The efficiency of a chromatographic method in the separation of the analytes and resolution of the peaks is described by the Van Deemter equation. In this equation, many different factors contributing to the retention time of the analytes and the peak broadening are taken in account (Van Deemter et al. 1956), although is not described here, because is beyond the aim of this thesis. The only parameter used in this work has been the logD. In facts, in the chromatographic method used, only the polarity of the mobile phase was changing, passing from Water 0.1% Formic acid to Methanol 0.1% Formic acid; with good approximation (±3 minutes) we can assume that the main parameter determining the retention time is the logD, as demonstrated by Creek et al. (2011), Boswell et al. (2011) and Silvester (2013). The logD is a parameter that can be calculate from the structure of compound, and some commercial software are present on the Internet and aloud the calculation of the theoretical logD ("ACD/Lab® 12.0 ChemSketch" and "ChemAxon® Marvinview 5.3.9").

3.2 Mass spectrometry

Mass spectrometry (MS) is an analytical chemistry technique that measures the mass to charge ratio (m/z) of the ions generated by the application of strong electric fields to compounds mixture. Ionization is the starting step of every mass spectrometric analysis, and can be achieved in many different ways. Every compound during ionization creates a specific ions cluster (spectrum); the type and intensities of the ions produced by a compound depends on the ionization technique and the instrument used. This means that a compound ionized under the same analytical conditions creates the same spectrum. This fact implies that comparing the spectra of the chemical standards to the spectra of the samples, whether two spectra matches at the same RT, the sample contains such analyte.

A mass spectrometer is an instrument consisting of three different parts: 1) ion source, 2) mass analyzer, 3) detector. Many different instruments exist and are used currently by mass spectrometers. A brief description of the most common instruments used in metabolomics will follow, with a deeper description for the ones used in this thesis.

Basic parameters used in mass spectrometry will be introduced here:

A) Sensitivity: detection of as many ions as possible. Higher sensitivity means more ions detected by the instrument

B) Mass accuracy: Accuracy in the determination of the m/z ratio of a compound.

C) Mass resolution: is the ability to distinguish two ions with slight difference in mass to charge ratio. It is calculated as $R = \frac{M}{\Delta M}$ where R = resolution, M = the m/z of the second ion, and ΔM is the difference in m/z of the two ions.

D) Molecular ion: "an ion that results from the loss of an electron by an organic molecule", according to "The free dictionary", http://encyclopedia2.thefreedictionary.com.

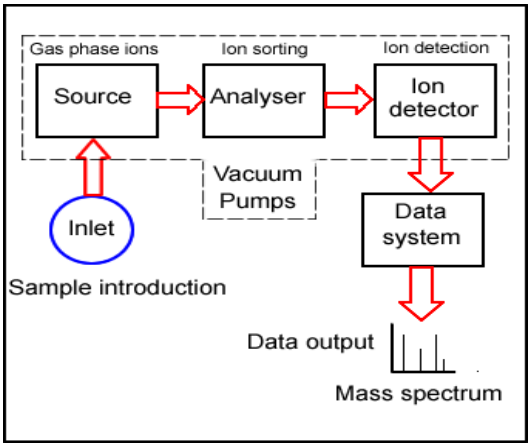


Image 3: A Schematic representation of a Mass spectrometer. The inlet introduces the sample in the source, where the analytes are ionized. Electric fields drive the ions through the mass analyzer where they are separated according to their m/z ratio. At the end of the path, a detector, detects the passage of ions. All the information are collected by a data system (computer), that sorts the data and gives a graphical representation of the signals collected by the mass spectrometer. Every part of the mass spectrometer is under vacuum, but the pressure in the distinct compartments is different.

3.2.1 Ion sources

Multiple ion sources exist in mass spectrometry, but only few of them are widely used and are available for metabolomics analysis. Here a list of the main sources used in metabolomics studies: electron ionization (EI), chemical Ionization (CI), electrospray ionization (ESI), atmospheric pressure chemical ionization (APCI).

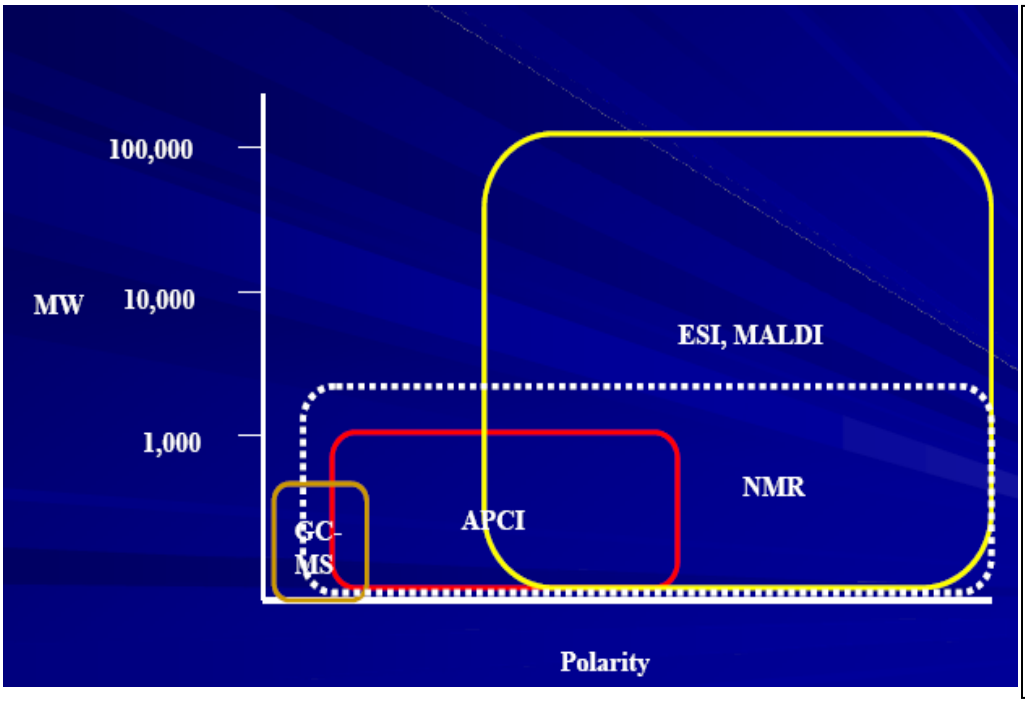


Image 3: In the image, the different kind of metabolites that is possible to analyze with the different techniques is displayed. The X axis is a theoretical scale of the polarity of the compounds, while in the Y axis the molecular weight (in Dalton) is displayed. The area of the compounds that is possible to study by NMR is also displayed; indeed it can be considered as complementary technology to MS.

3.2.1.1 Electron ionization

Electron ionization (EI) is an ionization technique used together with gas chromatography. It applies a strong electron beam to gaseous molecules that add or subtract electron from the gas phase according to the formula

$$M + e^- \rightarrow M^{+\bullet} + 2e^-,$$

The energy applied during the ionization is very high, 70 KeV; this maximizes the ionization of the gas phase molecules, but at the same time the high energy breaks some bonds in the molecular structures, creating numerous ionized fragments. The main advantage of this ionization technique is that the created spectrum is very similar between different instruments, and it is peculiar for every compound. A huge database of EI spectra exists (NIST mass spectral library), where is possible to compare the personal results obtained in the laboratories to many millions of molecules analyzed using EI ionization. The disadvantage of such technique is that the fragmentation is so strong that there is no signal for the molecular ion. Furthermore, strong fragmentation decreases the instrumental sensitivity. To overcome this problem, different sources are available for Gas chromatography, like CI (chemical ionization that will be treated next), APGC (atmospheric pressure gas chromatography) or GC-APCI (gas chromatography atmospheric pressure chemical ionization). The two latter did not become yet of standard use in metabolomics and will not be described in this thesis.

3.2.1.2 Chemical ionization

In chemical ionization (CI), the ionization is obtained using a gas (usually methane, isobutane or ammonia) that is in a larger amount in the source than the analytes. In the source, the gas is bombarded with electrons that charge the gas phase. The gas phase reacts with the analytes creating four different possible ions, according to the formulas

$$M + CH_5^+ \rightarrow CH_4 + [M + H]^+$$
$$AH + CH_3^+ \rightarrow CH_4 + A^+$$
$$M + C_2H_5^+ \rightarrow [M + C_2H_5]^+$$
$$A + CH_4^+ \rightarrow CH_4 + A^+$$

The advantage of this technique is that the fragmentation is soft, the pseudo-molecular ion is intact, while a simpler spectrum is produced. Chemical ionization is considered a soft ionization technique compared to EI, but its use is enclosed to few specific studies.

3.2.1.3 Electrospray ionization

Electrospray ionization (ESI) is the main ionization technique used in my thesis and will be described in its basic principles here (for a deeper description, ESI development has been reviewed by Ho et al. 2003). ESI is a soft ionization technique that was invented in late 80s and become the standard in liquid chromatography analysis. ESI is able to ionize liquid mixtures through the application of an electric field coupled with massive heating. It is the preferred choice for liquid chromatography although it is able to work with capillary electrophoresis and direct infusion (Dunn et al. 2013).

In ESI, the liquid phase containing the analytes is driven through a heated stainless steel capillary (at temperature between 100° to 150° centigrade) where an electric field from 1 to 5 kV is applied. Even if the process is not perfectly understood, it is assumed that the electric field passes electric charges to the liquid into the capillary. Whenever the liquid exits from the capillary, it encounters a chamber with higher temperature (between 250° to 500° centigrade); the pressure inside the capillary, due to the heat, creates a liquid spray. When each singular spray droplet enters the hotter chamber, it tends to evaporate faster, donating its charge to the remaining liquid molecules. As none of the analytes dissolved is evaporating faster than the liquid, analytes will receive the charge form the liquid droplets (Dole et al. 1968). At this point, the charged analytes are driven by electric fields to the mass analyzer. The addition of acids or salts to the mobile phase increases the conductivity, decreases the droplets size, therefore facilitate the evaporation (desolvation). Evaporation is a key factor; solvents that evaporate easily tend to give a better ionization. Conversely, water is not an optimal solvent due to its relatively high boiling point, and its direct injection limits ionization.

Many version of the electrospray ionization exist, for example nanospray, or cold-electrospray. Nevertheless, the original version is the most versatile and it is the preferred ionization source in metabolomics LC-MS analysis.

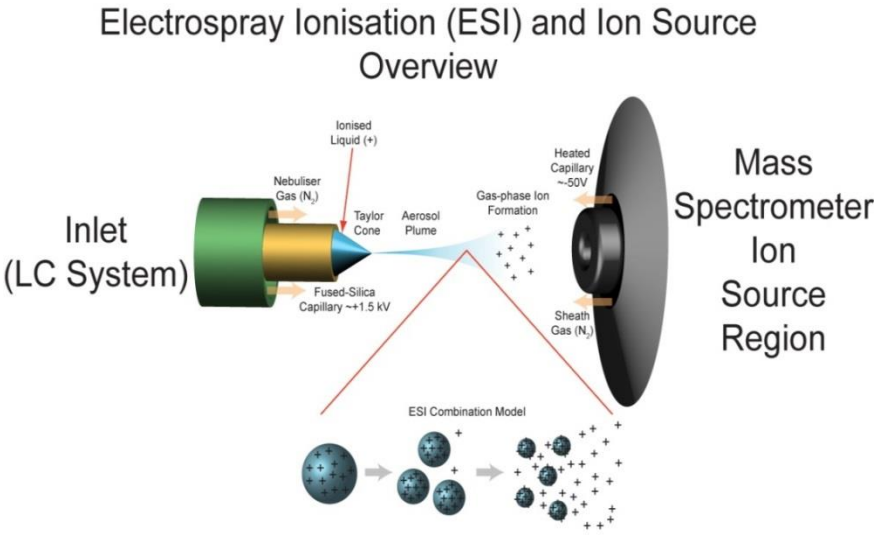


Image 3: A schematic representation of an ESI source. The capillary (yellow cilinder in the picture), gives a charge to the liquid solvent containing the analytes. The hot temperature and the pressure inside the capillary form a spray, where every droplet quickly evaporates, transferring its charge to the analytes. The ions enter the cone attracted by an electric field (cone voltage).

The ESI source is a soft-ionization device and it creates little fragmentation during ionization. It can be used in both positive and negative mode. In positive, it forms mostly pseudo-molecular ions $[M+H]^+$, but it can form also adducts with atoms present in the solvents, like $[M+Na]^+$ and $[M+K]^+$. In negative, it gives a pseudo-molecular ion of $[M-H]^+$, and adducts with chlorine $[M+Cl]^-$, or with the ionized forms of the acids present in the solvents, like Formic Acid = $[M+FA]^-$ or Acetic Acid = $[M+AA]^-$. Adducts with the solvents or with combined ions are also possible. Even fragments can ionize binding to adducts. Multiply charged ions are frequent in ESI; this phenomenon can be used to study entire proteins or their peptides.

Due to the multiple variables implicated, the ionization process varies between different sources and between the different solvents. Thus, it is not so reproducible, which means, for example, that every instrument from a different producer creates different spectrum for the same compound. This has been the main limit to the formation of extensive mass spectra libraries like the one existing for Electron ionization, even if some smaller libraries are now available, like Massbank (www.massbank.jp) and Metlin (http://metlin.scripps.edu).

3.2.1.4 Atmospheric pressure chemical ionization (APCI)

APCI is a complementary technique to ESI. Even if ESI is very versatile, less polar compounds do not ionize in ESI, while APCI is a better solution for non-polar metabolites.

APCI is similar to CI; usually it gives mono-ionized pseudo-molecular ions ($[M+H]^+$, $[M-H]^+$). Ionization occurs in two different steps: in the first step, the solvent coming from the LC evaporates very quickly due to a counter-current Nitrogen gas flow at maximum 600° centigrade. The vapor

droplets are then ionized by a corona pin discharger or a β-particles emitter. APCI ionizes the molecules in the gas phase, while in ESI the charge is given in the liquid phase, before spraying. This also allows using non-polar solvents, while in ESI it is not possible. It produces fewer fragments than ESI, and the fast desorption helps to prevent decomposition of the analytes in the source. Due to its structural similarity to an ESI source, the two sources are interchangeable (reviewed in Covey et al. 2008).

In this work, I tentatively used the APCI source to look at the ionization pattern of volatile compounds injected in liquid chromatography and to see if this technique can be applied to real samples. Even if the results were not encouraging, they will be discussed in chapter 5.

3.2.2 Mass analyzers

The mass analyzer is the instrumental part where the research has focused more during the last 30 years. This part of the instrument is dedicated to separation of the ions produced by the source, according to their mass-to-charge ratio (m/z). The theory behind the mass analyzers argues that applying the same force (usually electric or magnetic field) to multiple ions with the same charge, the acceleration of each singular ion depends only on their mass. This is stated in Newton's second law,

$$\mathbf{F} = m\mathbf{a}$$

Where F = force, m = mass, and a = acceleration.

To be more precise, the exact formula describing this process would be the Lorentz force law, which is the basic law of the motion of charged particles in vacuum.

$$\mathbf{F} = Q(\mathbf{E} + \mathbf{v} \times \mathbf{B})$$

F = force, Q = ion charge, E = electric field, v = ion velocity, B = magnetic field

Transforming **F** in **ma** and moving **Q** to the left side we obtain the classic equation of motion of the charged particles in vacuum:

$$(m/Q)\mathbf{a} = \mathbf{E} + \mathbf{v} \times \mathbf{B}.$$

If we use the elementary charge number **z** in the equation (z = Q/e), we obtain the measure of the mass to elementary charge ratio **m/z**.

The **m/z** can be measured in several ways: for example, as the spatial shift while an ion passes in a curved electromagnetic field (sector instrument); as the ability to pass a radio-frequency filter (quadrupole and ion trap instruments); as the time to pass through a drift tube (time of flight instruments); as the Fourier transform of the frequency of ions turning around electric or magnetic traps (Orbitrap or Ion Cyclotron Resonance instruments). In the next section, the mass analyzers applied in metabolomics studies will be described, focusing on the ones used in this thesis.

3.2.2.1 Quadrupole mass analyzers

Quadrupole mass analyzer consists of four parallel metal rods. Radio frequency (RF) voltage is applied to the rods, while a second current voltage is superimposed to the RF. Selecting the proper RF, only specific ions with a given m/z will be able to pass through the quadrupole, resonating inside the quadrupole rods. The ones with a different m/z than the selected one, will be deflected out of the path. Switching the RF voltage, it is possible to select ions with different m/z.

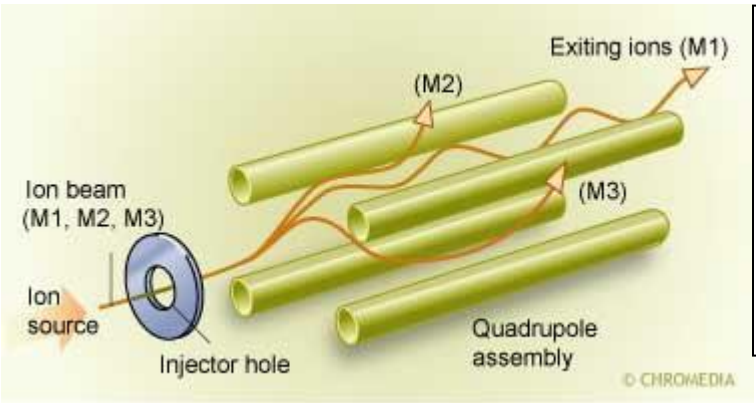


Image 4: An example of quadrupole. The ions coming from the ion source are driven through the quadrupole, where the are subject to an RF voltage. The ones able to resonate inside the quadrupole, go through the whole path, reaching the detector. The others are deflected against the quadrupole bars. Source: www.chromedia.org "Mass analyzers".

With quadrupole mass analyzer it is possible to select ions with a single m/z, or to scan in a given range continuously varying the RF voltage. The instrument works as a mass filter and is very fast. In the quadrupole, the selection of masses is dependent mostly on the rods position and on the applied RF voltage, so it needs little calibration. When a single m/z is selected the sensitivity for such ion is the maximum possible. On the other hand, the mass resolution of the instrument is low, because in a given RF multiple ion species may resonate and pass the quadrupole filter.

Due to the versatility of the quadrupole instrument it can be used in series like in triple quadrupoles (triple quad). In addition, combinations with different mass analyzers are possible, like quadrupole-time of flight (Q-TOF), or together with ion trap instruments. The triple quad and the Q-TOF have been used in my experiments, therefore they will be discussed in the MS/MS section of this chapter (section 3.3).

3.2.2.2 Time of flight mass analyzer

A Time of Flight mass analyzer (TOF) is an instrument that measures the time that an ion needs to fly from a starting point to the detector (Guilhaus, 1995). In the last generation of TOF mass analyzers, the ions produced from the source are forced by an electric field to fly from the starting point to the detector inside the flight tube. They enter the instrument in perpendicular way with respect to the flight tube, and a pulser pushes them inside. At around half of the flight tube, the ions are deflected at 180° (or more) by the reflectron and towards the detector. As the electric force applied is the same, the acceleration of the particles through the flight tube depends only on the mass to charge ratio. For example, if the source creates only single-charged-ions, the acceleration will depend only on the mass (this is the case for MALDI sources, while ESI tends to create also multiply charged ions).

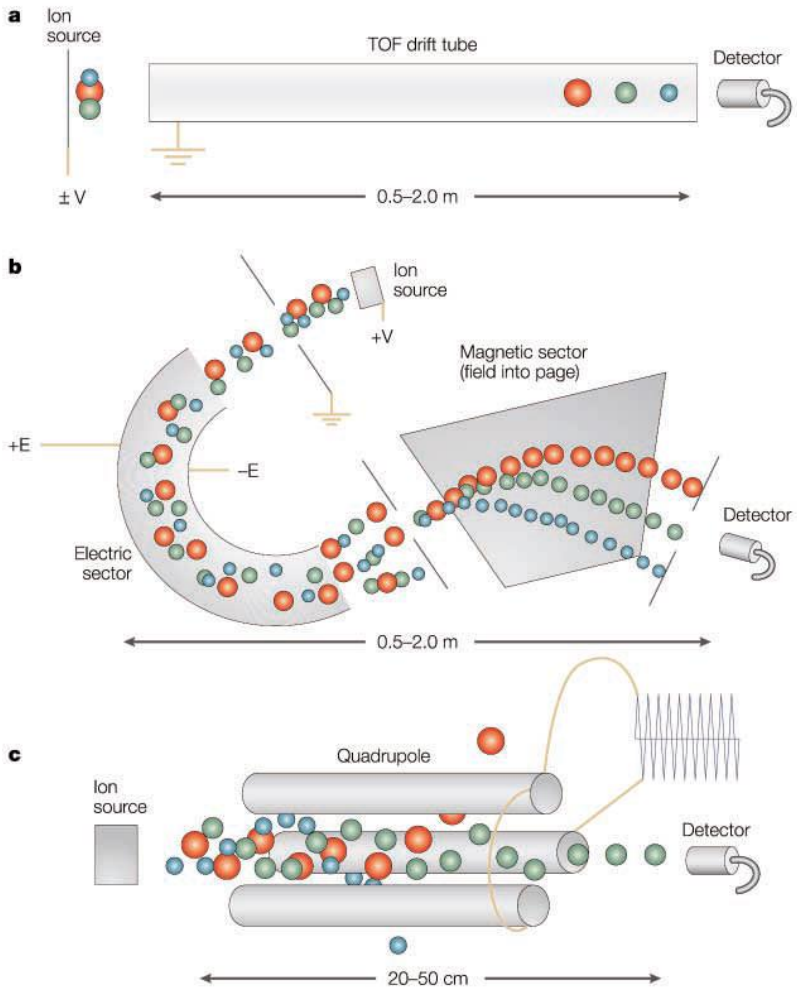


Image 5: A comparison between a TOF drift tube, an electromagnetic sector and a Quadrupole. In the TOF, ion separation is achieved by the different acceleration of the ions under the same electric field. In the electromagnetic sector, ion separation is achieved through a curved sector that gives a short path to lighter ions and a longer to heavier ones. In the quadrupole, a radio frequency determine the ions able to resonate inside the quadrupole path and the ones forced to walk out the path..

The electromagnetic sector mass analyzer is not described in the thesis, because is not used in metabolomics experiments. Nevertheless, also TOF instruments take advantage of the separation that a curved electromagnetic field can give to the ions. A part of the TOF, called Reflectron, deflects the ion of about 180° during their route inside the flight tube and improves ion separation in the TOF instrument. Better ion separation = Better resolution. Source: Glish and Wachet, (2003).

The ions produced in ESI source have a momentum and initial speed different from each other. While entering the TOF tube, they need to be focused in a small spatial position (1-2 millimeters). In order to achieve it, they collide with a residual inert gas in the RF guide. When the ions are all in the same spatial position, they are pushed through the TOF tube. The TOF tube is in perpendicular position in comparison to the initial direction of the ions. Starting the flight path with the same initial speed, the resolution is sharply increased in comparison to ions entering with different speeds.

The reflectron also increases sharply the resolution. It uses an electrostatic field to deflect the ions at 180°; the lighter ions switch their direction faster than the heavier ions. This increases the separation between ions of different m/z. The presence of a single reflectron is called V-mode, while also multiple reflectrons can be used to increase the difference between the ions. The latter setting is called W-mode.

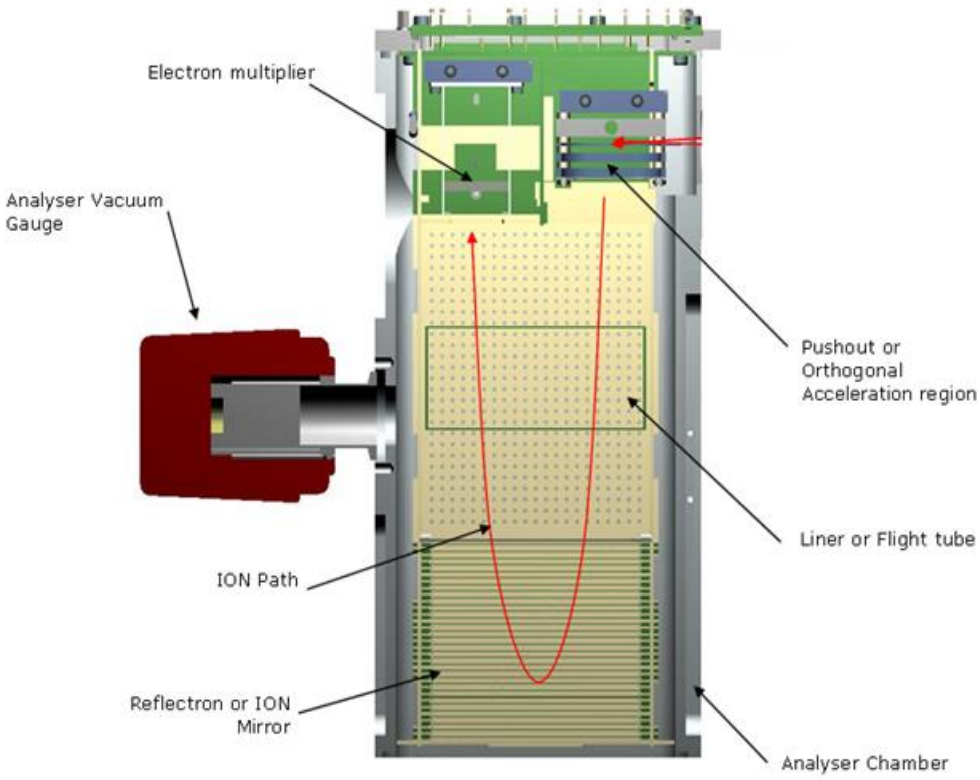


Image 6: A schematic representation of a V-mode TOF mass analyzer. The ions enter perpendicularly to the flight tube. They are focused in a specific position and then pushed through the tube. At half of the path, the reflectron, deflect the ions of about 180° and they reach the detector (an electron multiplier in this case) , where the signal is transmitted to the computer. The TOF is in extreme vacuum, because interfering molecules decrease the mass resolution.

First TOF patents appeared in the 50s, but the first prototypes were not so sensitive, and their dynamic range (ability of detect diverse orders of magnitude of ion current) was very low. TOF instruments had a very low diffusion up to the 90s, when the invention of new sources (ESI and MALDI) and new detectors, allowed the use of the TOF. Modern TOF instruments have a dynamic range of five orders of magnitude, maximal resolution of 50 thousands and sensitivity slightly lower than triple quad instruments. The application of new sources and new technologies, permitted the use of TOF instruments also with Gas chromatography and Capillary Electrophoresis. Nevertheless, their main application is coupled to Liquid Chromatography.

3.2.2.3 Fourier transform mass analyzers

Fourier transform mass analyzers are instruments that use the mathematical Fourier transform to convert the signal of the machine to m/z spectrum. Two instruments use this tool: FT-Orbitrap and FT-ICR-MS. Orbitrap is a trap mass analyzer able to trap the ions around its central fuse in elliptical trajectories in harmonic mote. It has been designed by Makarov in 2000 (Makarov et al. 2000) and its patent has been acquired by Thermo-Fisher scientific, that is the only producer and vendor in the world. The m/z measurement is based on the detection of the trajectories of the ions through the image current

induced to an outer electrode, which works as an electrical amplifier. To improve the resolution, the signal must be detected many times, so the Orbitrap cannot work in continuous mode. To inject the ions in the Orbitrap, a C-trap interface is used to trap and inject ion-packets in the Orbitrap. Then, the Fourier transform converts the signals to an m/z spectrum. The nominal max resolution of the newest version of the Orbitrap reaches now 280.000.

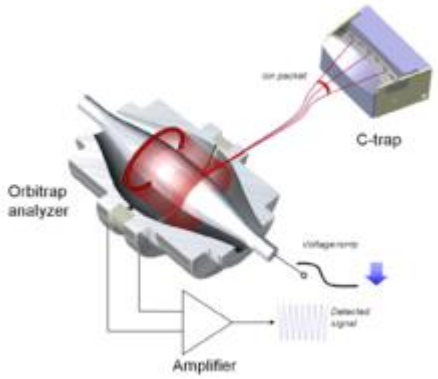


Image 7: Scheme of an Orbitrap mass analyzer. The ions are firstly trapped in the C-trap, and then are sent in packets to the Orbitrap. While entring, the voltage is decreased, up to let the ions reach the orbit, then increased to make them spin around the fuse. Turning around the fuse, the current transmitted by every singular ion is registered and amplified by the detector. The Fourier transform algorithm converts the signal to m/z. Source: Thermo Fisher Scientific.

FT-ICR-MS has a concept similar to the Orbitrap, whereas it is rather complex. In FT-ICR, the ions are introduced in the instrument applying a quadrupolar electric field that drives the ions towards the magnet. The perpendicular constant magnetic field makes them rotate in harmonic trajectories around the magnet, all together. If the magnetic field is constant, the ions rotate in constant cyclotron frequencies. These frequencies are measured by metal plates and the signal is back converted to m/z through Fourier transform, according to the equation:

$$\omega_c = \frac{qB}{m}$$

This equation is just an approximation to let the reader understand the relationship between the magnetic field and the m/z (Glish and Wachet, 2003). Increasing the magnetic field, the Cyclotron frequency of the ions increases as well. In higher frequencies, even ions with very similar m/z separate, increasing mass resolution and mass accuracy. This is the reason why, with very strong magnets, the ICR can achieve sub-ppm mass error and millions in mass resolution.

3.2.3 Detectors

The detector is the last instrument present in a Mass spectrometer and converts the signal of the ions to an electric signal readable by the computer station. Many different detectors exist, but they can be classified in two classes: the ones hit by the ions that multiply the signal, and the ones that measure the ion frequency (FT-MS instruments). In the first class there are Faraday's cups, Ions to photon detectors and electron multipliers (and its derivatives).

A "Faraday's cup" works as following: the ion hits the metal cup and the charge (electron) is passed to the metal where it is transmitted producing electricity. The amount of electricity produced is proportional to the amount of ions hitting the metal plate. The system is not very sensitive, as single ions produce electricity of one single charge (electron). However, the proportion between the signal and the amount of charges is very high, thus Faraday's cups are used to quantify the signals.

In the "ion to photo detectors" the ions hit a scintillator compound (a compound able to emit photons), and the compound emits photons that are then detected by a light detector. Usually the scintillator compound is interfaced with a microchannel plate; the ions hit the microchannel plate that releases electrons targeting the scintillator compound. The Ion-to-photo-detectors have good sensitivities, especially if interfaced with microchannel plates although they may result noisy.

The electron multipliers take advantage of the fact that an ion or electron hitting a secondary emissive material can liberate from one to three electrons. In the electron multipliers, the emitted electrons are further accelerated by an electric field to emit more electrons. At the end, from one single charge, thousands of electrons are emitted. Microchannel plate detectors are simply multiple Electron multipliers that are spatially divided. They can also provide spatial resolution, for a better measurement of the m/z.

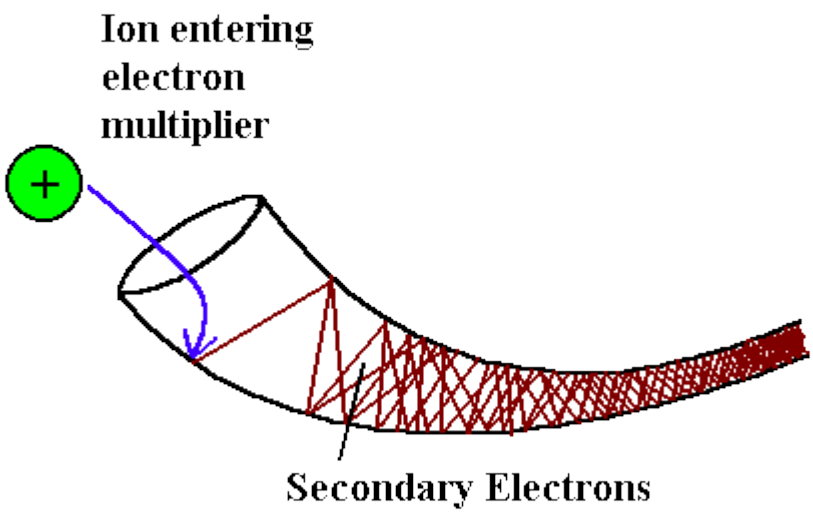


Image 8: an electron multiplier. The ions enter in the electron multiplier and hit the surface. The surface releases from 1 to 3 electrons that re-hit the surface that again releases from 1 to 3 electrons. At the end a big amount of electrons are released from the instrument. The current produced is directly correlated to the amount of ions entering in the detector. Source: http://www.webapps.cee.vt.edu/ewr/environmental/teach/smprimer/icpms/icpms.ht

For a wider description of all the detectors developed for mass spectrometric analysis, I suggest the review of Neetu et al. (2012).

3.3 Tandem mass spectrometry (MS/MS analysis)

Tandem mass spectrometry consists of two or more mass analyzers, both performing measurements during the same analysis to improve identification and/or quantification of the analytes. This combination of mass analyzers is often called "Hybrid MS". The tandem MS/MS analysis is a generic name of a process in which an ion formed in an ion source is mass-selected in the first stage of analysis, reacted, and then the charged products from the reaction are analyzed in the second stage of analysis (Glish et al. 2008).

Many combinations of different mass analyzers are possible, and in general, the smaller and simpler mass analyzers (quadrupoles and ion traps) can easily be coupled with any of the other mass analyzers; the same holds true with slightly different versions of themselves. As the subject is very vast, I will describe only the "hybrid" mass spectrometers used in this work, referring to the literature for the other existing and future hybrid possibilities.

3.3.1 Triple quadrupole (QqQ)

A triple quad (QqQ in this case), is an instrument where three different quadrupoles are in series one after the other. In the common scheme, the first (Q_1) and the third (Q_3) quadrupoles work as mass filters, like a singular quadrupole would do, scanning or selecting singular ions according to the RF voltage applied to their rods. The second quadrupole, instead, is set to act as a collision cell, where a gas (usually Helium, Argon or Nitrogen) is present, and an electric field is applied. The electric field excites the gas, increasing exponentially the probability of accidents between the gas molecules and the analytes ions. Because of the "crashes", the analytes structure collapses, forming fragments that are successively measured by the third quadrupole. This kind of fragmentation is called CID (collision-induced dissociation).

Summarizing, the first quadrupole either selects or scans the ions coming from the source, the second quadrupole generates fragments from the ions filtered by the first quadrupole, and the third quadrupole selects or scans the fragments obtained. Different combinations of the selection or scanning in the first and third quadrupole allow performing different experiments.

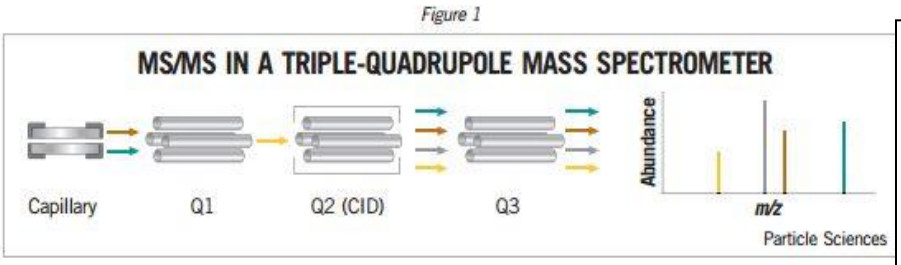


Image 9: A schematic representation of the triple quad. On the right is displayed a typical MS/MS spectra in Q3 scan mode. Source: http://www.particlesciences.com/news/technical-briefs/2009/mass-spectrometry-bioanalysis.html

In my thesis, I only used the LC-ESI-QqQ-MS instrument with the MRM method (multiple reactions monitoring) to precisely quantify some metabolites (described in chapter 7). In the MRM method, the first quadrupole selects only the ion of interest, discarding all the other ones; the second quadrupole is set to perform the best fragmentation for such ion, to obtain few relevant fragments. The third quadrupole is set to select only the expected fragments, discarding the remaining ions. In one MRM experiment, usually at least three ions are selected: one in the first quadrupole and two in the third quadrupole. The one higher in intensity between the two detected in the Q_3 is called quantifier, because it is directly correlated with the one detected in the Q_1 and it is used to quantify the signal. The other one is called the "qualifier" because its detection serves only as further evidence that the selected analyte is the right one and is producing the expected fragments.

When analyzing the data, the computer shows a peak, if and only if, all the selected ions are generated in the mass spectrometer. The peak area is directly correlated with the concentration of the analytes in the sample. The quantification is generally done creating a calibration curve with multiple injections under the same conditions of the chemical compound in analysis with different dilutions. If the signal falls in the linear range of the instrument, a linear calibration curve is generated and it is used to quantify the signal peak obtained from the sample analysis at the same retention time, giving the absolute concentration of the metabolite in the sample.

3.3.2 Quadrupole-Time of Flight-Mass spectrometer (Q-TOF-MS)

A Q-TOF is an hybrid mass spectrometer that takes advantage of the selectivity of the quadrupole and the scanning properties of the TOF, in order to improve the mass resolution (in MS mode) and to fragment selected ions to be scanned in the TOF mass analyzer (MS/MS mode). This setting has also a collision cell that often is a quadrupole, either a Hexapole or an Octapole. Despite the different number of rods, their function does not differ from a quadrupole, they are selected only because at the same RF, an Octapole can transport a wider mass range of ions in comparison to a quadrupole.

In the MS/MS mode, the quadrupole is used as a mass filter to select ions of a specific m/z. In the collision cell, the selected ions are fragmented through CID and analyzed in the TOF sector. The use of the TOF allows having high-resolution MS spectra and high-resolution MS/MS spectra; high-resolution spectra help in compound identification. On the other hand, in MS/MS mode, during the fragmentation, the impacts decrease differentially the ions' speeds, generating an ion beam delayed in time and space. The focus of the fragment ions in the TOF pusher is lower than in the MS mode, so the resolution and the mass accuracy will be lower. Sometimes, because the quadrupole selects the ions with a mass range of about one Dalton, it is possible that parent ions with a slightly different m/z might be selected together for the same MS/MS analysis, and in the collision cells fragments from both parent ions are produced. This effect confuses the MS/MS spectrum obtained in the Q-TOF. The interpretation of such spectrum may be not possible.

In my work, I used a "Synapt G1 HDMS" (WATERS, Manchester UK, Image 10), that is an Electrospray Ionization-Quadrupole-Ion Mobility Shift-Time of Flight-Mass Spectrometer (ESI-Q-IMS-TOF-MS). In SYNAPT the ions produced in the source are guided through a Z-shaped path, to eliminate the un-ionized molecules that entered the path by chance. The ions are then driven to the quadrupole, where they are scanned (in MS mode) or selected (in MS/MS mode). After the quadrupole, the ions are driven to the TOF, where they are pushed in the flight tube in the orthogonal direction. There are two operative modes for the TOF: the V mode, where ions are deflected by the Reflectron directly to the detector, and the W mode, where the ions are deflected three times from two different Reflectrons before reaching the detector. While the V mode has higher sensitivity and a resolution of about 10000, the W mode has a lower sensitivity and a resolution of about 17500 with an increased mass accuracy. In MS analysis, I used the W mode, to have a clear measurement of the ion masses, while in the MS/MS analysis I used the V mode, to detect the major number of fragments possible. In MS/MS mode, the ions were fragmented in the "transfer", using a CID fragmentation similar to the one described in the section 3.3.1.

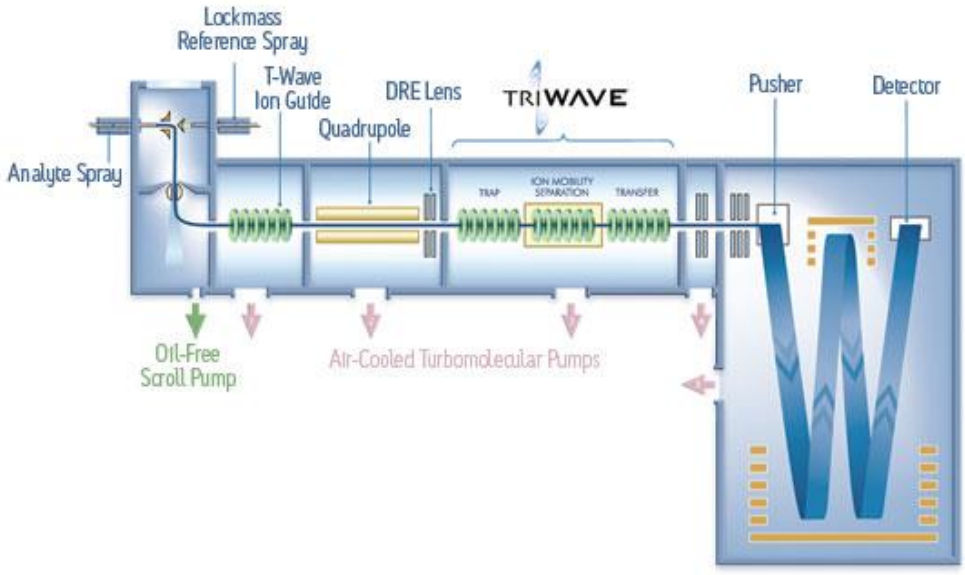


Image 10: The Synapt HDMS instrument used in my thesis. In the Synapt, the ions produced from the source are guided to the quadrupole where they are scanned. After passing the triwave (IMS), the ions are guided to the TOF, where the flight tube separates the ions and the time to flight from the entrance to the detector is registered. In MS/MS mode, the ions are selected in the quadrupole and fragmented in the transfer, before entering the TOF. Source: www.waters.com

As shown in image 10, this instrument has also a third mass analyzer, the Ion Mobility Shift (IMS), which is a very promising technology, orthogonal to RT and to m/z measurement. In basic principles, IMS separates the ions according to their structural shape, forcing them to pass through a tridimensional grid formed through an inert gas that delays in time ions with different shapes. In this work I am not going to describe further the IMS because I did not use it, due to the fact that the version installed on this machine is the first launched on the market by WATERS, and it was not suitable for Metabolomics analysis (Franceschi et al. 2011). Its limit was that it could not cover all the whole mass range of 50 to 1500 Dalton used in my experiments (50 to 1500 Dalton is the common mass range for metabolomics analysis). A wider description of the IMS is demanded to literature (Kanu et al. 2008).

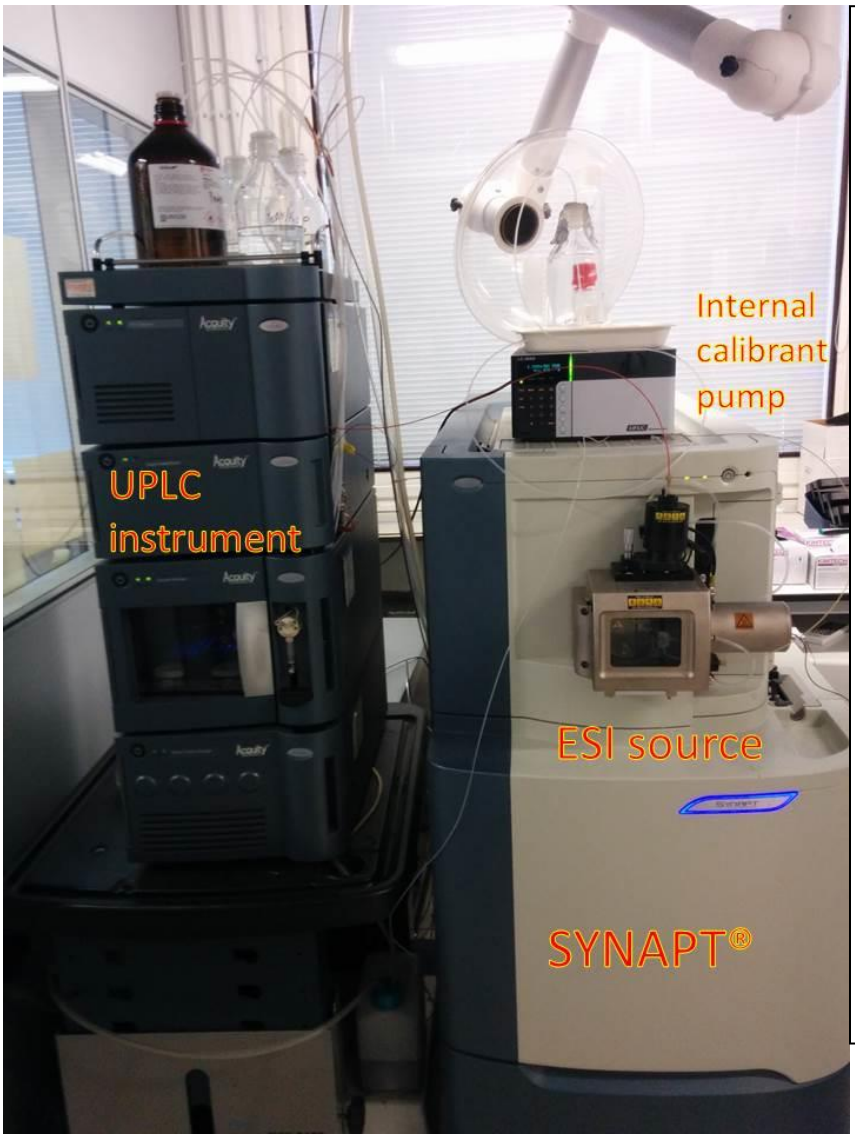


Image 11: The UPLC-SYNAPT (UHPLC-ESI-Q-IMS-TOF-MS) instrument used in my experiments. The UPLC is directly connected to the ESI source of the mass spectrometer. Both UV-vis spectra and MS spectra were acquired during the analysis. UV-vis was scarcely used in my experiments, and data is not shown. As you can notice from the picture, also a second pump is connected to the ESI source: indeed internal calibrant Leucine-enkephaline 10ug/L is pumped in continuous flow of 0.1 ml/min in the instrument. The resulting m/z 556.2780 (in positive) and 554.2615 (in negative) were used to perform on-line calibration of the acquired spectra. The use of the internal calibrant has be proven to be a very effective method to maintain the calibration of the TOF stable for days. TOF calibration is very sensitive, especially to temperature, because temperature shift, affects the kinetic energy of the ions in the flight tube, increasing or decreasing their flight time. Leucine-enkephaline calibration can cope to temperature change up to ± 4° centigrade.

.

References Chapter 3

1. *Arapitsas, P., Speri, G., Angeli, A., Perenzoni, D., & Mattivi, F. (2014). The influence of storage on the "chemical age" of red wines. Metabolomics, 10(5), 816–832. doi:10.1007/s11306-014-0638-x*
2. *Boswell, P. G., Schellenberg, J. R., Carr, P. W., Cohen, J. D., & Hegeman, A. D. (2011). A study on retention "projection" as a supplementary means for compound identification by liquid chromatography-mass spectrometry capable of predicting retention with different gradients, flow rates, and instruments. Journal of Chromatography A, 1218(38), 6732–41. doi:10.1016/j.chroma.2011.07.105*
3. *Covey, T. R., Thomson, B. A., Schneider, B. B., & Introduction, I. (2009). Atmospheric pressure ion sources, (October 2008), 870–897. doi:10.1002/mas*
4. *Creek, D. J., Jankevics, A., Breitling, R., Watson, D. G., Barrett, M. P., & Burgess, K. E. V. (2011). Identification by Retention Time Prediction, 8703–8710.*
5. *Dettmer, K., Aronov, P. a, & Hammock, B. D. (2007). Mass spectrometry-based metabolomics. Mass Spectrometry Reviews, 26(1), 51–78. doi:10.1002/mas.20108*
6. *Dole M, Mack LL, Hines RL, Mobley RC, Ferguson LD, Alice MB (1968). "Molecular Beams of Macroions". Journal of Chemical Physics 49 (5): 2240–2249. Bibcode:1968 J. Ch. Ph. 49. 2240D. (doi:10.1063/1.1670391)*
7. *Dunn, W. B., Erban, A., Weber, R. J. M., Creek, D. J., Brown, M., Breitling, R., … Viant, M. R. (2013). Mass appeal: Metabolite identification in mass spectrometry-focused untargeted metabolomics. Metabolomics, 9(SUPPL.1), 44–66. doi:10.1007/s11306-012-0434-4*
8. *Franceschi, P., Vrhovsek, U., & Guella, G. (2011). Ion mobility mass spectrometric investigation of ellagitannins and their non-covalent aggregates. Rapid Communications in Mass Spectrometry : RCM, 25(7), 827–33. doi:10.1002/rcm.4932*
9. *Gika, H. G., Theodoridis, G. a., Vrhovsek, U., & Mattivi, F. (2012). Quantitative profiling of polar primary metabolites using hydrophilic interaction ultrahigh performance liquid chromatography-tandem mass spectrometry. Journal of Chromatography A, 1259, 121–127. doi:10.1016/j.chroma.2012.02.010*
10. *Glish, G. L., & Vachet, R. W. (2003). The basics of mass spectrometry in the twenty-first century. Nature Reviews. Drug Discovery, 2(2), 140–50. doi:10.1038/nrd1011*

11. *Glish, G. L., & Burinsky, D. J. (2008). Hybrid Mass Spectrometers for Tandem Mass Spectrometry. Journal of the American Society for Mass Spectrometry, 19(2), 161–172. doi:10.1016/j.jasms.2007.11.013*
12. *Guilhaus, M. (1995). Principles and instrumentation in time-of-flight mass spectrometry: Physical and instrumental concepts. Journal of Mass Spectrometry, 30(September), 1519–1532. doi:10.1002/jms.1190301102*
13. *Ho, C. S., Lam, C. W. K., Chan, M. H. M., Cheung, R. C. K., Law, L. K., Lit, L. C. W., … Tai, H. L. (2003). Electrospray Ionisation Mass Spectrometry : Principles and Clinical Applications, 24(February), 3–12.*
14. *Kanu, A. B., Dwivedi, P., Tam, M., Matz, L., & Jr, H. H. H. (2008). SPECIAL FEATURE: Ion mobility – mass spectrometry, 1–22. (doi:10.1002/jms)*
15. *Makarov A. (2000). "Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis". Analytical Chemistry : AC 72 (6): 1156–62. (doi:10.1021/ac991131p)*
16. *Neetu, K., Ankit, G., Ruchi, T., Ajay, B., & Prashant, B. (2012). A review on mass spectrometry detectors. International Research Journal of Pharmacy, 3(10), 33–42.*
17. *Silvester, S. (2013). Mobile phase pH and organic modifier in reversed-phase LC–ESI-MS bioanalytical methods: assessment of sensitivity, chromatography and correlation of retention time with in silico logD predictions. Bioanalysis, Vol. 5, No. 22 , Pages 2753-2770 (doi: 10.4155/bio.13.250)*
18. *Theodoridis, G., Gika, H., Franceschi, P., Caputi, L., Arapitsas, P., Scholz, M., … Mattivi, F. (2012). LC-MS based global metabolite profiling of grapes: solvent extraction protocol optimisation. Metabolomics, 8(2), 175–185. doi:10.1007/s11306-011-0298-z*
19. *van Deemter JJ, Zuiderweg FJ and Klinkenberg A (1956). "Longitudinal diffusion and resistance to mass transfer as causes of non-ideality in chromatography". Chem. Eng. Sc. 5: 271–289. doi:10.1016/0009-2509(56)80003-1*
20. *Vuckovic, D. (2012). Current trends and challenges in sample preparation for global metabolomics using liquid chromatography-mass spectrometry. Analytical and Bioanalytical Chemistry, 403, 1523–1548. doi:10.1007/s00216-012-6039-y*

4. Metabolomics: the basic concept, experimental design and data analysis

The definition of Metabolomics is not an easy task and this word is mostly unknown to the public. According to the Oxford dictionary, Metabolomics is "The scientific study of the set of metabolites present within an organism, cell, or tissue". This means that Metabolomics aims to individuate what metabolites and how much of them are present in an organism, cell and tissue, under specific conditions. In a typical metabolomics experiment two groups are compared: the first group (group A) as a control, and the second group (group B) as a treatment. The metabolomics analysis aims to find the metabolites that have different concentration between the group A and the group B. Often metabolomics results are coupled with previous knowledge (Physiology, Genetics or Morphology) and/or other omics sciences (Transcriptomics, Proteomics and Genomics) to give a physiological understanding of the differences found between A and B.

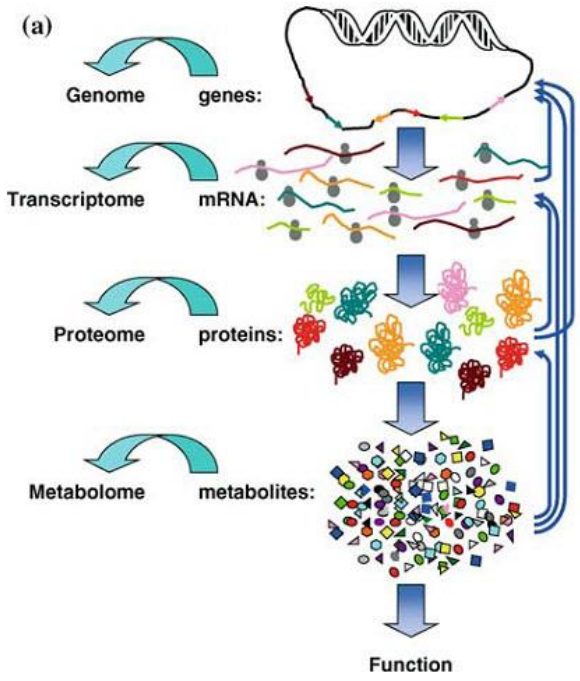


Image 1: The relationship between the four main "OMIC" sciences. The whole set of the genes (genome), transcribe to the whole set of the transcripts (Transcriptome) that are translated to the whole set of the proteins (Proteome), which determine the presence and amount of the whole set of metabolites (Metabolome). Source: http://schaechter.asmblog.org/schaechter/2009/05/of-terms-in-biology-metabolomics-and-metabonomics.html

In MS-based metabolomics, two different kinds of analysis are possible, targeted and untargeted, and the choice of one or the other determines a different experimental approach, in both sample preparation and data analysis (reviewed by Patti et al. 2013). In targeted MS metabolomics analysis, the compounds that are going to be measured in the samples are previously chosen (targeted)based on their availability on the market, their detectability on the LC-MS instrument and their importance in the

experimental design. To detect the chosen compounds, their injection as pure standards under the same instrumental conditions is performed before and during the analysis, to determine their detectability and to create a calibration curve that will be used to give a precise quantitation of each metabolite. The chosen compounds are often representative of few specific metabolic pathways, already known to be present in the biological sample under analysis. During the analysis, only the targeted compounds are going to be detected. The preferred instruments to perform this kind of analysis are "triple quads", due to their high sensitivity, robustness, and precision in compound quantification. This analysis allows describing to which extent there is a difference in concentration of the given metabolites between the group A and the group B. It allows seeking whether any of the given metabolites shows a clear different pattern between the two sample sets. When enough known compounds are present in the method, it is possible to describe the metabolic pathways regulation and to infer about the physiological status of the two sample sets.

In untargeted MS metabolomics analysis, there is no previous choice of the compounds measured in the experiment. The goal of this experiment is to collect all the signals (ions) obtained from the samples, comparing two sample sets and finding which signals are significantly different between the two sets. Each ion generated from a metabolite is part of a spectrum of ions obtained from the same metabolite. The data are acquired by instruments able to collect spectra for a wide mass range(Q-TOF, Orbitraps). The diverse ions obtained in different sample chromatograms are first aligned sample by sample, grouped in pseudo-molecular spectra and then analyzed statistically to find the differences between sample sets. The resulting biomarkers need to be validated and then identified. The last step is the inference on the biological meaning of the results.

The two analyses have very different approach to the data: in the targeted analysis, the measurement of the metabolites is limited - only tens to few hundreds can be measured in the same analysis (generally no more than 150). Their detection and quantification is compared to the relative pure chemical standard previously injected under the same analytical conditions, therefore a concentration curve is built to provide a precise (absolute) quantification of the metabolites in the sample set. The data analysis is then straightforward, because statistical inference is achievable through common univariate methods; the experimental design is not complex, and it can be easily adequate to the needs of the experiment. In the untargeted analysis, the number of the measured metabolites is unknown, the number of signals collected is very high, and their quantification is relative. The data analysis in this case requires many steps, and it needs to be carefully handled. The success of this kind of experiment is determined by the experimental design that needs to take in account the possible

variability present in the samples. Because of the huge amount of data obtained in a single experiment, the data analysis is the main bottleneck of this approach.

In my thesis I used untargeted MS metabolomics analysis. The approach to this methodology will be extensively described in this chapter. To perform an experiment involving untargeted MS Metabolomics analysis multiple steps are required to achieve interpretable data, listed here:

1) Adequate experimental design
2) Untargeted analysis: instrumental requirements.
3) Data analysis: pre-processing
4) Statistical analysis
5) Compound identification: spectral matching and putative identification.
6) Data mining
7) Data sharing

The seven step listed here will be developed in this chapter of my thesis, both from the theoretical point of view and describing the application that I have been developing in my experiments.

4.1 Adequate experimental design

As mentioned earlier, the experimental design is a very important step in untargeted MS analysis. Hundreds of metabolites can be measured in an untargeted analysis, and each metabolite can generate up to tens of ions in the MS-source, for a total amount of more than ten thousands of signals. Due to the high amount of variables (signals) detected in a single experiment, multiple factors can influence the outcome and create troubles in the statistical analysis. A first step to improve the quality of the data is to select the adequate number of samples (observations) per group, to allow a clear statistical result. The ratio variables/observations is extremely unfavorable to the statistical tests, and the risk of finding false positives and false negatives arises with the increase of the variables/observations ratio. Moreover, the average accuracy of the untargeted data is usually lower in respect to well-designed targeted methods, since the calibration of unknown compounds cannot be finely tuned metabolite by metabolite. The classical approach (3 vs 3 observations) is not the right choice in untargeted MS experiments. In theory, higher the number of observations, more reliable the result of the statistical tests.

On the other hand, the samples need to be analyzed all together in short periods, to avoid that technical/instrumental variability and sample degradation may influence the outcome of the analysis (systematic variability). The factors influencing the results can be multiple: the operator (different

operators may prepare the samples in slightly different ways), the stability of the instrument (stability of the stationary phase and stability of the MS instrument, especially the cleanliness of the source) and the stability of the samples (degradation processes might happen during the analysis). Other minor factors can also influence the outcome: slightly different preparation in the mobile phase, different atmospheric conditions (in particular fluctuation of the temperature is challenging) during analysis and others. Nevertheless, this may happen in every experiment and precautions to prevent these problems are part of the good laboratory practices, so they will not be discussed here.

The goal is to A) prevent all possible biases and B) provide indicators that are able to signal when something goes wrong. A) Preventing all the biases due to poor experimental design or sample preparation and analysis is a key step to achieve good results. First, the adequate number of samples need to be selected. This number ought to be maximum possible according to these parameters:

1. Availability of the materials: Many times, the availability of the materials is limited, especially for the group of samples under treatment. The golden rule is to analyze all the samples under treatment and to keep a comparable number of samples for the controls. Indeed, selecting a very different number of samples for the two groups, increases the false positives in the group over-represented, and gives a higher number of false negatives in the group under-represented.
2. Preparation times: samples prepared in different days might show an inter-day variability, due to the fact that the conditions of extraction are mutated (different temperatures, humidity, operators). The stability of the extraction method needs to be proved and validated, injecting multiple times the same sample and assuring that the analytical outcome of such sample is the same, day after day.
3. Analysis time: the stability of the instruments is a key parameter. Usually the analytical method needs to be validated and the stability time needs to be assured, to understand when the instrument may need cleaning and resetting of the initial conditions.
4. Consider the expected difference between the two groups in order to estimate the power of the experiment: if the control and the treatment are very different, few samples are enough to extrapolate the right biomarkers, while if the difference is very small, hundreds of samples might be necessary to avoid false negatives.

As example, I will describe the experimental design of my own experiment (chapter 7):

1. 14 different grape varieties have been comparatively studied. I collected three different clusters per each grape variety for a total of 42 clusters. From each grape berry, I manually separated three main tissues (skin, pulp and seeds). Out of the 42 samples, 21 were representative of my "control" group (*vinifera* grapes), 12 were representative of the "treatment" group (American grape species) and nine were hybrid grapes that have been kept out of the statistical analysis. The limiting factor was, in my case, the number of ripening American grapes. Only four of them reached the correct ripening stage, so the selection of the number of samples was due to this parameter.
2. The diverse berry tissues have been separated before extraction. As the comparison was intra-tissues, I have extracted every tissue from all the samples the same day. The extraction method has been previously validated (Theodoridis et al. 2012).
3. The samples from all the tissues have been randomized and injected in the instrument in random order, both in positive and negative ionization mode. The LC-MS method has been previously validated (Arapitsas et al. 2013).
4. The number of samples depended on the number of ripening American grapes. This might have been a big limit in my analysis. Nevertheless, the expected difference between the samples was high, because I was comparing different species. On the other hand, comparing different species might be a limit, because their ripening stage could be very distinct between the samples. The maturation of the grape is the main parameter to consider when measuring the concentration of different metabolites: Sugars level increases during veraison (Coombe & McCarthy, 2000), while acids level decreases; tannins decrease in amount and became softer at taste, while aroma precursors are accumulated during ripening (Robinson and Davies, 2000, Conde et al. 2007). Therefore, the goal is to collect the grapes when they have the most similar maturation stage possible.

   In grape physiology, there are some methods to establish grape maturity. In my experiment, I used the measurement of the soluble solids (brix°) in the grape juice, because it is the easiest, quickest and it has been used for decades. This measurement is performed with a refractometer and this value is directly correlated with the sugar content of the grape; often it is coupled with the measurement of the acidity to evaluate the organic acids level. I collected the grapes when they reached a level of 18° brix.

A second method would have been to collect the grapes when the seeds turn brown; the change of the color is an indicator of maturity and edibility of the grapes. This method is destructive of the samples, and since I had very few grape berries for some of the American *Vitis*, it was discarded. A third option would have been to collect the grapes at the same "days after full blooming" (DAFB) or "days after pollination" (DAP). As my samples were located in fields and I needed numerous different berries, from different species (this is also a very limiting factor, because different species have very different maturation process), the measurement of these parameters was practically infeasible and this method was discarded.

B) The need of indicators able to signal whenever something goes wrong is necessary in every experiment. In metabolomics multiple types of indicators can be used to eye-catch eventual drifts.

1. Blanks: as every laboratory practice, a negative control is required to detect the presence of false signals.
2. Standard mixes (STDmix): injection of a mixture of known standards every 10 to 20 sample analyses may indicate if there is any drift in RT, MS calibration or peak intensities. Pure standards are very different from sample matrix, therefore STDmix samples are not considered a good indicator for the quality of the analysis.
3. Internal standards: samples can be spiked with standards not commonly present in the matrix (radiolabeled standards or non-common metabolites). They can be very useful to evaluate extraction efficiency, although their use as normalizer for data analysis is not suggested, because theoretically we would need to have a standard for each chemical class, across the whole RT. Too many internal standards would be necessary.
4. Quality control samples (QCs): QCs are a mixture of the extracted samples under analysis. They are injected every 6 to 10 samples, and their stability can be evaluated both with EIC of single metabolites and with Multivariate data analysis, as described in section 4.4. Their use is recommended by Gika et al. (2012) and Dunn et al. (2012).

4.2 Untargeted analysis: instrumental requirements.

Untargeted LC-MS metabolomics analysis requires the contemporaneous measurement of hundreds of metabolites. To aloud the identification and quantification of the metabolites, the instrument should be able to obtain separated signals for each metabolite. Technically, this is translated

in chromatographic and mass spectrometric peak resolution. Resolution is the ability of an instrument to separate two close-by peaks. Higher the ability, higher the instrumental resolution.

In chromatography, the chromatographic resolution is the ability of having two separated peaks closely eluting from the column. It can be calculated by the formula:

$$R_s = (T_{R2} - T_{R1}) / ((0.5 * (w1 + w2)) \quad (1)$$

Where the *R* is the resolution, *T* is the time of the two peaks and *w* is the width at half peak (Snyder et al. 2010). Higher the resolution, higher the capability of the instrument to obtain separated peaks. In the last years, many improvements have been done in the separation techniques, especially with the introduction of UHPLC (ultra-high-pressure-liquid-chromatography) which uses smaller particle size to improve peak resolution and with the commercialization of numerous columns that might be used at high temperatures and wide pH range. On the other hand, the perfect separation technology does not still exist; indeed it is not possible to perfectly separate metabolites from multiple chemical classes, so Metabolomics LC-MS method are built taking in account the highest number of known metabolites and the best separation possible.

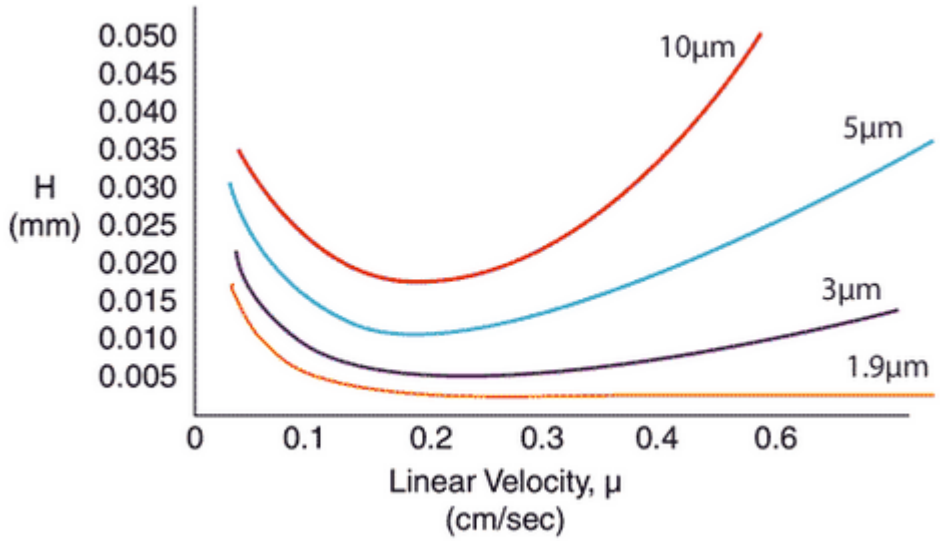


Image 2: The difference in H (height of the theoretical plate) between different column particle size. A lower H value indicates an higher number of theoretical plates per length of the column. Column with the same length will have an higher number of theoretical plates when their particle size is smaller.

In LC-MS based metabolomics, instruments with very high MS resolution are needed to overcome the limitation given by the limited chromatographic separation. Indeed, if the metabolites are not separated chromatographically, their separation can be achieved from a mass spectrometric point of view. The theory behind mass resolution has been described previously in chapter 3. The importance of the mass resolution is somehow the same of the chromatographic resolution. The goal is to

distinguish signals coming from diverse metabolites and to be able to identify each of them. Only high-resolution instruments can be successfully used in untargeted analysis, TOF, Orbitrap and FT-ICR. TOF instruments can have a resolution of 5.000 up to 60.000. Modern Orbitrap instruments can achieve up to 280.000 of resolution, while FT-ICR can reach up to 10.000.000 (Knolhoff et al. 2014).

When the resolution is higher than 100.000, the isotopes with the same nominal mass can separate according to their elemental composition. This happens because the mass defect of the isotopes of a compound can be different depending on the distribution of the isotopes in their formula. Therefore, the isotope of a compound having a Carbon 13 in its formula would have the same nominal mass but a different mass defect from the isotope having an Oxygen 17. With a resolution of over 100.000, is possible to see two distinct signals for the two distinct isotopes. When resolution is lower than 100.000, the instrument measures the different isotopes as a unique signal, averaging their mass defect accordingly to their content in the molecular formula. Even if it looks trivial, in presence of high mass accuracy, the averaged value can be used to improve formula calculations (Thurman & Ferrer, 2010).

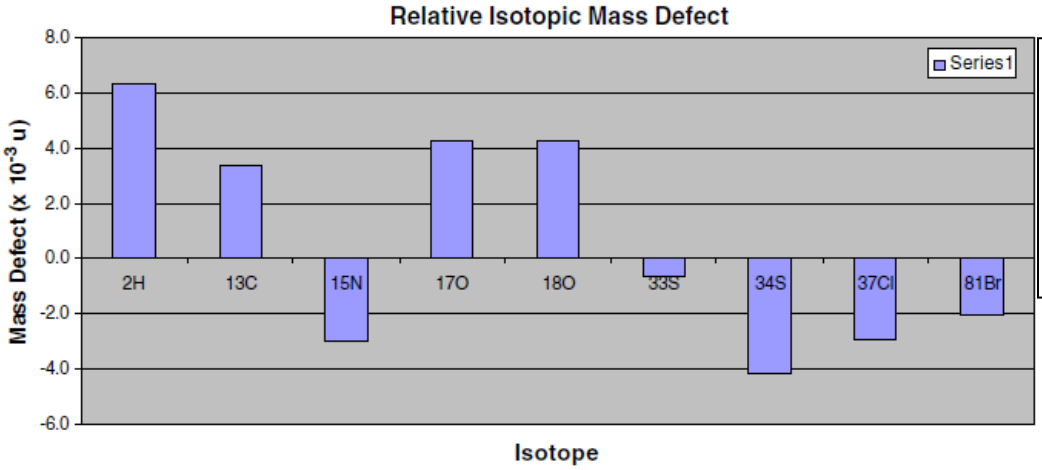
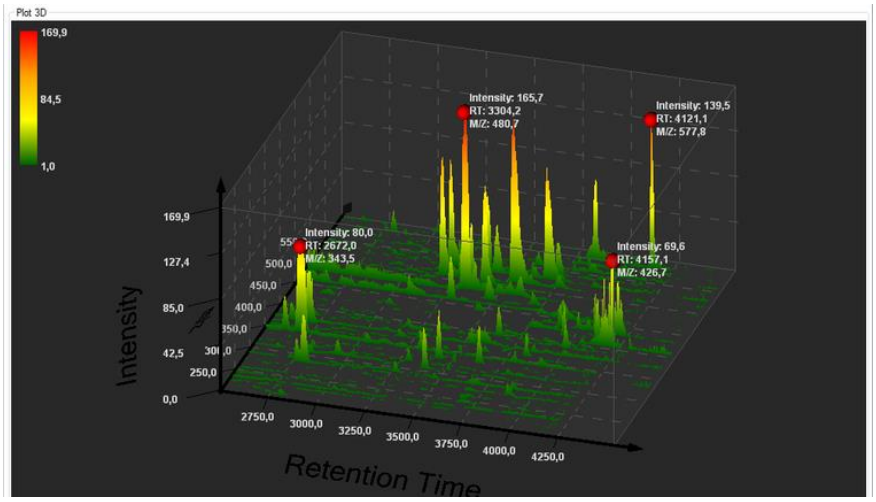


Image 3: The relative isotopic mass defect of some elements. Picture from the publication of Thurman & Ferrer, (2010).

4.3 Data analysis: pre-processing

After the instrument acquires the data, this is stored as chromatographic raw data files (one per sample) containing the information about the m/z detected, their retention time, and their signal intensity. To aloud a comparison between the different samples, the data need to be pre-processed to enclose all the information in a readable two-dimensional peaktable. This step is crucial for a correct interpretation of the data.

The first step is the *peak picking;* it tries to allocate all the chromatographic peaks per each m/z value. Their area is then integrated and it is (in theory) proportional to the corresponding compound concentration in the samples. Therefore, every integrated peak is defined by its m/z and RT and it is called *feature.* Peak picking needs to be performed on every sample and then the list of mz/RT features from a sample needs to be merged to the ones from all the other samples.



mz	rt	F001_D	F002_S	F003_C	F004_C	F005_C
753.193	9.74437	0	1.46453	8.67436	14.4593	4.33721
375.992	9.74516	2.76941	88.5246	5.86431	2.91132	2.93346
93.0331	9.7452	6.89842	4948.81	36.3696	29.3392	14.6666
1009.72	9.74779	0	0	91.4248	49.545	9.63176
1009.22	9.74839	0	0	186.158	127.364	23.2857
393.115	9.74961	0	4.07813	19.3357	23.2906	13.0982
1458.31	9.74961	0	1.38395	15.4467	30.7128	1.39575
433.012	9.75073	5.56179	32.8778	13.1947	8.88046	16.2775
313.093	9.75083	1.39699	20.6453	53.6023	106.716	33.316
314.094	9.75383	1.40463	1.38395	13.9528	15.3564	12.5617
755.165	9.75433	0	1.36991	63.0551	305.364	33.0232
267.987	9.75701	2.74274	121.566	8.57025	10.0046	11.4315

Image 4: A graphical representation of the peak picking process. In this process, the tridimensional information obtained from the chromatograms is transformed to bidimensional tables of features. Every features has is unique m/z, RT and intensity value per each sample. Due to the errors that are possible in peak elution and peak mass measurement from the LC-MS instruments, the picked peaks need to be aligned, grouped and eventually filled by an automatic binning step.

In theory, it is enough to sub-divide the whole range of m/z in bins obtaining the EIC per each bin, integrate it in every sample, and then group the features detected. This approach has been demonstrated to be error-prone, because of the instrumental error and the noise signal that is always present in LC-MS analysis. In facts, the chromatographic peaks are often miss-aligned and the MS detection suffers of error, which might be of several ppms in the different instruments (Shahaf et al. 2013, Knolhoff et al. 2014). For example, in the instrument used in my experiment, the average error is

around 5 ppms, but it is likely to have errors up to 30 ppms, and sometimes it can reach up to 100 ppms (Shahaf et al. 2013).

Many software have been released to perform peak picking: In my case, the vendor (Waters) supplied us with "Markerlynx" (Frederiksen, 2011), an automatic software for peak detection. Nevertheless, the outcome of such software was not directly usable with other tools for statistical analysis and peak identification, so I decided not to use it. In our laboratory, an automated pipeline for peak picking, alignment, gap-filling and pseudo-spectra construction has been developed by Franceschi et al. (2014). The pipeline is based on XCMS (Smith et al. 2006) and CAMERA (Kuhl et al. 2012). To perform the analysis with such software, the raw data files need to be first converted to NetCDF files using the *Dbridge* software supplied by Waters. The NetCDF files are then peak picked using XCMS with settings dedicated to the instrument used.

After peak picking, the next important step is to match the ions produced by the same metabolites across samples. This is called "alignment". XCMS uses the matched features to create a correction model for the retention time, to correct eventual RT shifts across samples. Last step is the filling-peak process: during peak picking, the peaks are detected above a given threshold; this would result, in a table with some zero values where the intensity was below the given threshold. For this reason, XCMS uses the detected features as bins to collect the EIC of the bins in the samples with zero values. After the fill-peaks function, the XCMS software gives to the user a so-called "peaktable", containing the values m/z, RT and intensity for every feature detected. This table do not contain information regarding the relationships between the different features. Indeed the co-eluting features might be originated from the same metabolite.

CAMERA software is dedicated to the discovery of the various relationships between the detected features. First, it divides the "peaktable" in RT bins, then it tries to relate the features each other through correlation analysis, and finally it calculates the various relationships between the features according to the m/z differences between them. Isotopic patterns and adducts are detected and labeled in the "peaktable". In the pipeline developed by Franceschi et al. (2014), the last step of the analysis is the automatic detection and matching of features corresponding to compounds already analyzed in the same instrument under the same conditions. At the end of the analysis, the peaktable contains all the detected features, their grouping, their putative relationships (adducts/isotopes) and the automatically identified compounds labeled with their "ChemSpider" code and putative names.

	ChemSpid	compoun	pcgroup	adduct	isotopes	mz	rt	C001_BLA	C002_STD	C003_QC_	C004_QC_	C005_MO
2961			43		[195][M+1	328.0978	1.376571	1.394262	4.183038	80.2519	92.17293	56.98027
2962			43			456.1747	1.377907	1.40336	0	21.05179	20.16111	12.63355
2963			43		[105][M+1	268.078	1.38292	0	4.183038	54.35266	44.11579	64.15155
2964	388379	raffinose	43		[1263][M+	1076.34	1.383073	1.40336	1.403268	9.82417	16.84143	2.807457
2965	389538 57	melibiose	43		[47][M+1]	180.0632	1.383172	4.182784	6.970885	148.6967	136.5314	114.9796
2966			43			185.0224	1.383272	4.210081	9.822876	181.3252	162.5066	238.7587
2967			43		[914][M+2	651.202	1.383535	1.40336	1.403268	30.09486	33.54954	18.24847
2968	389538 60	melibiose	43		[1][M]+	59.01321	1.38383	0	4.160155	1093.734	1155.509	1001.73
2969			43			172.0594	1.384158	5.546252	2.773437	38.83031	42.79009	48.40503
2970			43			312.9805	1.384261	8.167087	255.6671	31.30976	24.5043	27.22701
2971	5805	xylose	43		[468][M]+	457.1235	1.386387	0	1.403268	534.7124	460.6828	243.9246
2972			43			372.106	1.386883	0	0	21.62187	27.83869	29.03224
2973	7149 3895	D-(+)-treh	43		[221][M+2	343.1194	1.387239	1.40336	1.403268	414.9518	344.6433	419.6771
2974	6019 5576	D-(+)-malt	43		[592][M]+	503.1668	1.387636	4.210081	7.01634	353.5387	285.8806	371.1757
2975			43		[1107][M+	832.2813	1.387636	1.40336	0	19.58765	22.45524	9.826098
2976			43		[1088][M+	800.2439	1.387659	1.413975	2.827746	12.7276	1.414379	7.15193
2977			43		[1101][M]	815.2845	1.388146	1.40336	0	132.3378	117.5691	68.78269
2978			43		[1107][M]	831.2859	1.388146	0	0	25.26214	36.48977	14.95138
2979			43		[239][M+1	354.078	1.388239	2.773126	1.386718	132.5237	167.9728	139.595
2980			43		[63][M+1]	206.0409	1.388282	4.182784	5.577046	40.24559	56.59335	40.17547
2981			43			251.0534	1.388413	2.788523	2.788523	131.7679	114.3502	64.43285
2982	389538 60	melibiose	43		[221][M]+	341.1119	1.388655	44.96334	32.33477	9202.034	10597.96	11252.44

Image 5: an example of the peaktable outcome from the pipeline published by Franceschi et al. (2014). In the example the first columns represent respectively the chemspider code for the putatively assigned compounds, their names, the CAMERA group, the possible adducts and isotopes.

4.4 Statistical analysis

4.4.1 Systematic variation assessment and data normalization.

The first step to perform statistical analysis is to evaluate if there is presence of systematic variation, due to the instrument or to the experimental design. To assess the systematic variation some methods exist. The classical approach used in every MS experiment is to have technical replicates; the variability within the technical replicates needs to be lower than the 25% of the %CV, to assume that no systematic variation or drifts are present in the analysis. This method is very reliable, but in practice is not applicable to the untargeted analysis, because the number of variables to consider in the comparison is too high. A compromise of this method is to spike every sample with "internal standards", a mixture of compounds that are not naturally present in the samples under analysis (Sysi-Aho et al. 2007). The spiked compounds are then evaluated for their percentage CV with a threshold of 25%, as reported above. The problem of this approach is that we would need an internal standard representative per each chemical class present in the sample; this means that we need several internal standards. In my experiment I used 3 different internal standards (3-Indole-propionic acid, 4-stilbenol and Gentisic acid), but their use in the stability assessment has been secondary in comparison to the assessment based on the quality control samples (QCs) evaluation method developed in our lab.

The approach applied in our laboratory is based on the multivariate evaluation of quality control samples. Quality control samples (QCs) are a mixture of the various samples present in the

experimental design that were injected several times in the LC-MS system during the analysis. Their chromatograms were evaluated like normal samples, extracted with the pipeline described in the section 4.3 and the data visualized in a PCA plot in the first two components (PCA will be described in the following section). The stability of the analysis is assured by the fact that the QCs need to be all plotting in the same area, without diverging around in the PCA plot. As the QCs has been injected across the entire sample list, if they plot all together, most likely the systematic variation is null or very low. Otherwise, the presence of QCs in different positions means that analytical drifts may have occurred during the analysis. The 25% percentage of CV threshold can be coupled with the multivariate approach, using some known metabolites from the QCs. This method is very convenient especially because this analysis – using the described pipeline - can be directly performed during data acquisition: it is enough to convert the up-to-date acquired data in NetCDF files, to submit the files to the pipeline and to perform PCA statistical analysis, to see whether drifts are occurring. Since systematic variations, due to the operator, materials, and extraction method, have been previously avoided, drifts are more likely from the LC-MS instrument. A restore of the initial analytical conditions of the instrument (like LC cleaning, MS source cleaning and calibration) is often more than enough to restore the initial performance and continue with the data acquisition (Arapitsas et al. 2012).

This method, if performed correctly, eliminates the necessity of normalization. In the experiment performed in chapter 7, I did not perform any data normalization; indeed, due to experimental constrains, normalization has been used in the experiment described in chapter 5. Normalization methods are comparatively revised by De Livera et al. (2012).

4.4.2 Multivariate statistical analysis

Metabolomics data is multivariate by its nature, namely because multiple variables are often correlating, and may have both linear and non-linear distributions. Metabolomics data is often analyzed using multivariate statistical analysis (MVA). Many MVA methods exist; here I will briefly describe only the more common ones, focusing on their use in metabolomics.

The most common MVA in metabolomics is Principal components analysis (commonly called PCA; for a wider description on PCA I suggest the tutorial of Jolliffe, 2002). PCA aims to individuate and extract the main patterns in a dataset. A linear combination of the correlating features is extracted in a given number of so-called components, ordered according to the amount of variability explained by each component (the components are calculated according to the NIPALS algorithm developed by Herman Wold in 1975). This means that the first component always includes a higher amount of

variability than the second component, the second more variability than the third, and so on. In the image 6, an example of PCA plot. As you can notice, the first component (on the x-axis) represents more than 50% of the variability, while the second component (on the y-axis) less than 10%. As the goal was to distinguish the American *Vitis* grapes from *Vitis vinifera* ones, the plot clearly shows a separation between the two groups along the first component. In the green set, there are five red dots indicating the plotting of the QCs. As shown, they plot very close each other, meaning that there was not a systematic variation across injections. In the second plot on the right, it is possible to read the names of some samples. In my experiments I had three "technical replicates" (three different berry clusters per grape variety), and in the PCA they plot closer than other samples, indicating that reproducibility of the analysis was high.

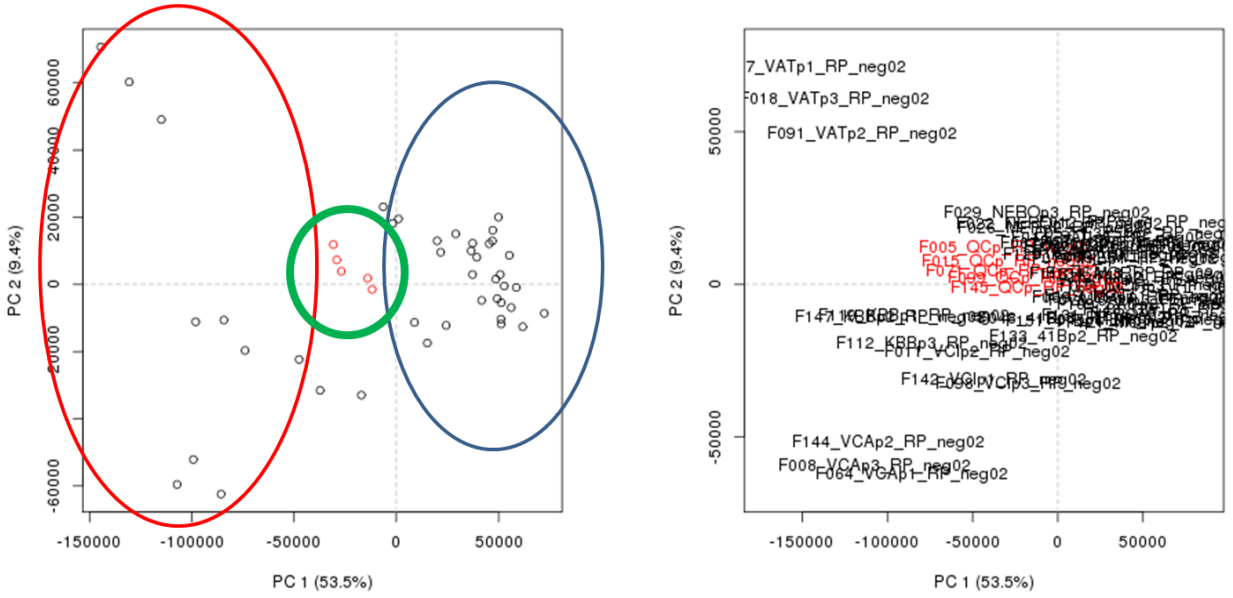


Image 6: On the left, the PCA plot of the injections of the grape skins of all the samples plus the QCs (in red). QCs plotting all together indicate lack of systematic drift during the analysis. The true groups (red and blue set) were separating along the first component which included the highest amount of variability. On the right, the same PCA plot with the names of the samples.

Nevertheless, the scope of this work was not to demonstrate that the American *Vitis* berries are different from the *Vitis vinifera* berries, but to find the variables that determine this difference. As in PCA plot the two groups were not uniform (the samples were from different species and different varieties), I needed a statistical method to focus only on the difference between American vs *vinifera.*

In MVA metabolomics, the common practice is the use of Orthogonal Projection to latent structures discriminant analysis (O-PLSda) also called Orthogonal Partial Least Squares discriminant analysis (Bylesio et al. 2007, Rosipal 2007). In O-PLSda, the user defines the categories of the samples that are stored as a matrix of zero and one (the so-called Y matrix, where the samples assigned to 0 are not part of such category, and to 1 if they are part of it) and the variability orthogonal to the categories will be excluded. The original PLS algorithm finds a linear regression model between the variables matrix (X matrix) and the categories matrix (Y matrix) projecting them into new spaces. The point is to find the multidimensional direction of the X matrix that explains better the multidimensional directions in the Y matrix. Modern PLS algorithm are also able to take in account the non-linear relationship between X and Y matrices (Wold et al. 1989, Wold et al. 2001).

The main difference between PLS-da and PCA is that the former tries to fit the components focusing on the variables, which determine the highest difference between the two chosen groups. It can be represented as a shift in the hyper plane, determining a focus on the space where the highest distance between the two groups exist (image 7). If the goal of the PCA is to group the variability of the X matrix in components, in PLS-da the goal is to focus the main difference between the groups into the first component. Its second component rather describes the variability intra-groups.

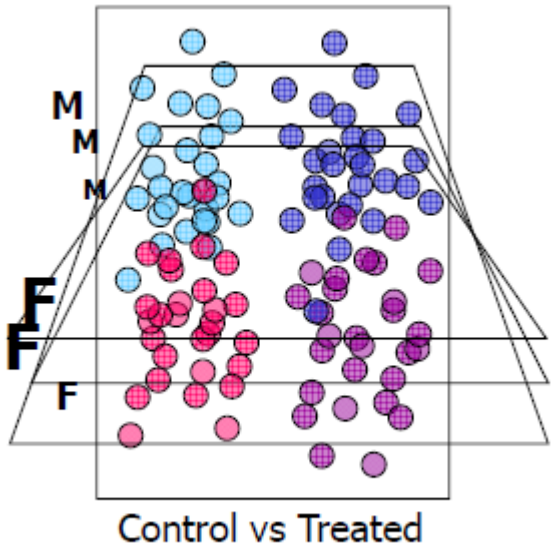


Image 7: An explanatory scheme of the plane shifting happening between OPLS-da and PCA. The goal of OPLS-da is to group in the first component all the variables that determine the highest difference between the two groups. This image is from the *Umetrics* training course on SIMCA-P+ 12.0.0

Once the PLS-da focused on the variability between the groups, it is possible to extrapolate the variables that determine such variability (biomarkers). The PLS-da loadings would be enough to extrapolate such data, but a more accurate harvest of the biomarkers can be achieved using the S-plot.

The S-plot is a graphical representation of the loadings according to their correlation and covariance; plotted on the bases of their significance in the groups. Marker variables are at the extremes of the plots, while non-significant variables are plotting in the center of the plot (Image 8). In this plot, there is not a threshold level, because the significance of the variables is dependent on the data and on the number of samples.

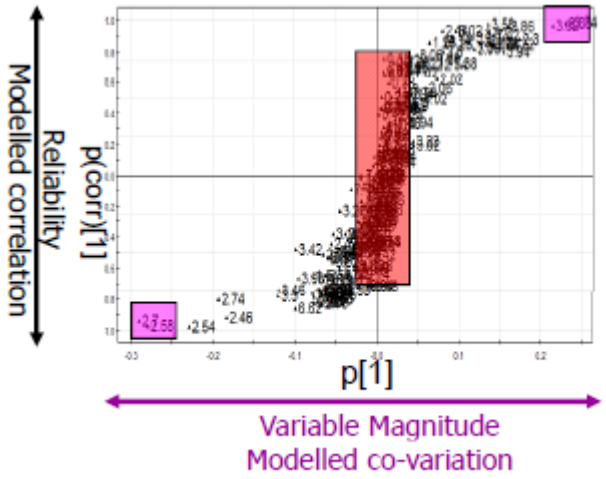


Image 8: S-plot of some spectrometric data. The horizontal axis indicates the covariance of the variables, while the vertical axis indicates the correlation of the variables across the samples. This image is from the Umetrics training course on SIMCA-P+ 12.0.0

4.4.3 Univariate statistical analysis

Despite of the multivariate nature of the metabolomics data, the use of easier univariate methods is possible, at some expenses. The number of variables taken in accounts needs to be limited as much as possible to increase the power of the test, and to decrease the false positives and negatives. A good method to limit the variables is to select only the ones overcoming a selected threshold (for example MS/MS experimental threshold), and the ones not respecting the 80% rule. After this step, univariate tests can be successfully applied to the data. A key point is to select the adequate test to analyze the data. In table 1 I summarize the test choice according to data pairing and distribution.

Experimental design	Normal distribution	Far from normal-curve
	Compare Means	Compare Medians
Compare two unpaired groups	Unpaired t-test	Mann-Whitney
Compare two paired groups	Paired t-test	Wilcoxon signed-rank
Compare more than two unmatched groups	One-way ANOVA with multiple comparison	Kruskal-Wallis
Compare more than two matched groups	Repeated-measures ANOVA	Friedman

Table 1: the univariate tests suggested according to the data structure. The table was taken from the work of Vinaixa et al. (2012)

In my experiment I chose the Wilcoxon t-test (Wilcoxon, 1945), because the distribution of the data was expected to be not normal for most of the metabolites, due to the distinct nature of the samples (especially in the American group different species have been considered as part of the same group). The selected p-value was 0.05. Although this is considered a good approximation in many experiments, a threshold of 0.05 means that a 5% of variables are expected to be false positives, which in my case of thousands of variables corresponds to a huge number of false positives. A numerous correcting methods can be used to overcome this problem, like Bonferroni, Holm, Hockenberg and Šidák (reviewed in Feise, 2002). Nevertheless, all these approaches become too stringent when a high number of variables is introduced in the test, increasing the risk of producing false negatives. The best choice in this cases is the so-called "false discovery rate" FDR (Benjamini & Hochberg, 1995). After the univariate test, the distribution of the variables in the range 0 to 0.05 of the p-value (the positive variables) is over-represented if compared to the rest of the distribution (negative variables), and this over-representation is expected to be the "true positives". This is then used to calculate a new cutoff value called q-value, which is the result of the following equation:

$$\frac{p - value}{positive\ variables}$$

This means that when I used a cutoff of 0.05 and I obtained 500 positive variables, my new cutoff (q-value) is 0.05/500= 0.0001. This is then applied to the original data. Applying the FDR, it is assumed that a 0.05 of the positive variables are allowed to be false positive ones. This method is less stringent than the correction methods and can be easily used with metabolomics data (Vinaixa et al. 2012). An example of the use of the univariate methods is given here, reporting the statistical analysis performed in chapter 7 (Image 9). The corrected p-values of the numerous markers can be coupled with the fold change parameter, which is calculated as the difference in folds between medians of the two

groups. Coupling these two parameters, it is possible to obtain a so-called volcano plot, where the variables are plotted according to their fold change (on the x-axis) and to the corrected p-value (on the y-axis). Common fold change thresholds are considered two or three. In my experiment, I decided to select 10 (one order of magnitude difference between compounds accumulation), to be able to focus on the compounds that are systematically diverse between the two groups and to exclude the differences due only to morphological or physiological status of the berries. Indeed, the berry shape, the distinct texture of the tissues across the samples and the physiological status might affect the accumulation and detection of the metabolites, with two or three fold changes due only to these differences. A very high threshold has been chosen to avoid intrinsic differences between the materials.

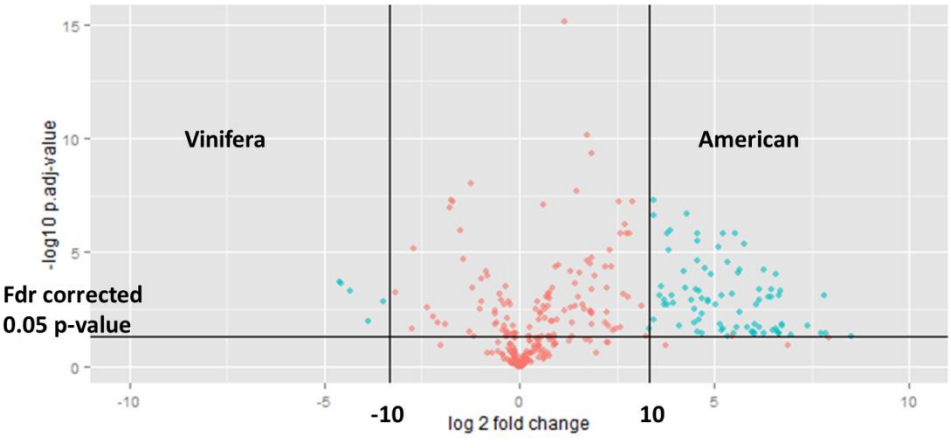


Image 9: A volcano plot, with the comparison between the *vinifera* berry fleshes and American berry fleshes. The red dots represent the variables below the threshold selected for the corrected p-values (0.05) and the fold change (10). The blue dots are the positive variables considered as putative biomarkers.

4.5 Compound identification: spectral matching and putative identification

Compound identification is considered the bottleneck in untargeted metabolomics experiments. To face the discussion is necessary first to delimit the identification concept. According to the metabolomics society initiative (MSI) there are four levels of identification (Sumner et al. 2007):

1. MSI level 1: compound identification through the direct comparison of at least two orthogonal characteristics with a reference standard.
2. MSI level 2: putative compound identification through spectral matching with a reference spectrum from spectral database.
3. MSI level 3: compound classification in a unique chemical class through spectral comparison with in-silico simulators.
4. MSI level 4: unassigned compound (only mz/RT signature).

The goal of the identification process is simply to move upward in the present scale. The outcome of the data processing is a peaktable consisting of thousands of features. Each feature has an mz/RT signature and the measured intensity across the samples. When also CAMERA analysis is performed, they are grouped according to their RTs and Pearson correlation of their area, known relationships between the features are labeled (isotopes and adducts). In Franceschi's pipeline there is also an automatic matching of the compounds with the in-house-build standards database present in our lab. The automatic matching is an important step for the identification of MSI level 1; indeed, the method uses "two orthogonal characteristics", like the retention time and at least two-reference m/z, and it compares them to reference standards. Even if the automatic matching needs a manual check to confirm its findings, it speeds up the identification process.

However, the remaining number of unidentified features is extremely high. The identification of all the compounds present in the data is beyond the metabolomics analysis, which aims only to individuate all the metabolites that differ between two groups of samples. Therefore, the aim of the analysis is to focus on the identification of the biomarkers. The statistical analysis delimits the importance of the variables to a restricted number of features. The next step used in metabolomics experiment is to individuate the "known" biomarkers, the features that can be easily identified comparing their spectrum to the one of a reference standard. In my experiments, Franceschi's pipeline identifications speed up the individuation of the "known" biomarkers.

The remaining un-matching biomarkers are labeled as unknowns and go to further identification steps. The identification path is generally composed of these steps:

1. Acquisition of the MS/MS or MS^n spectra and comparison of the spectra with external spectra databases or MS/MS spectra simulators.
2. Determination of the chemical formula based on the mass accuracy and isotopic pattern (Kind & Fiehn, 2007) and eventually MS/MS data.
3. Retention time prediction of the suggested structures through the construction of regression models using the calculated physico-chemical properties of such compounds (Creek et al. 2011, Bosweel et al. 2011), or through the direct comparison of the retention times with another external RT database where the suggested structures have been already injected (Stanstrup & Vrhovsek, 2014).
4. Biological meaning of the suggested compound structures: compound structures similar to metabolites already present in the matrix under analysis (or part of their metabolic

pathway), are more likely to be present in such matrix than compounds with no-biological evidence.

Nowadays, all the steps here described can be performed with computer-assisted strategies, with open source software like R. A deeper description of this strategy is given by Stanstrup et al. (2013), who has been my reference during data analysis. Indeed, a similar strategy with some modifications has been performed in the experiment described in chapter 7, and it will be described there. On the other hand, every step of this strategy has some limitations, listed here:

1. Acquisition of clear MS/MS spectra is not straightforward, especially in hybrid TOF instruments where multiple ions may be selected as precursors before fragmentation, falsifying the spectral interpretation. This problem is very limiting when ions with similar m/z are co-eluting. Poor MS/MS spectrum means no identification.
2. Calculation of the chemical formula sometimes might become not feasible. Poor isotopic pattern assessment (>5% of error in isotopic ratios is troubling), scarce mass accuracy (>5 ppm error is already a problem) and the presence of multiple elements in the formula, make this task very complex and sometimes unfeasible (Kind & Fiehn, 2006). For example, TOF instruments have been demonstrated to acquired data with up to 30 ppms error (Shahaf et al. 2013), while Orbitrap instruments are reported to have problems in the measurement of the second and third isotopes intensities, giving a misleading isotopic pattern ratio.
3. Prediction of retention times through regression model has been demonstrated to be reliable but not very helpful, because the range of RT predicted by the model is too wide to exclude all the structures similar to the right one. On the other hand, the use of direct comparison with the retention times from another chromatographic method has been proved to work a lot better, but it requires that the suggested compound has been already injected in another LC system.
4. The biological meaning of the data is not usable when none of the compounds of the same biosynthetic class has been analyzed previously in the same matrix with the same instruments.

In the near future, new instrumentation will probably solve the problems and break the limits here listed. Instruments with higher mass accuracy, better estimation of the isotopic patterns and higher chromatographic resolution are already in the market and will become of common use in the next

years. From my point of view, a different approach could be used and it has been developed in this thesis in chapter 6.

4.6 Data mining

Data mining is the last step of a metabolomics analysis. It differs from all the other steps because it is experiment-dependent, and it is different across the distinct experiments, according to the goals. The idea is to use the data obtained from the analysis, to understand what the physiological differences between the two subsets are. So, the goal of the data mining is decided during the experimental design and the means to achieve this goal depend on the final aim of the study.

To mine the data, statistical analysis is necessary. Statistical analysis performed before the identification can be coupled with further analysis to infer on the biological meaning. Usually, Pearson's correlation between metabolites, hierarchical clustering between samples (Everitt, 1993) or Bayesian inference on the metabolites relationships (Suvitaival et al. 2014, McGeachie et al. 2014) are performed to determine which of the identified biomarkers are related to each other and infer on their inner relationship. Graphical representation of the statistical relationships between metabolites can be done through Heatmaps, or using dedicated graphical software like Cytoscape (www.cytoscape.org).

On the other hand, the statistical analysis is not sufficient alone; random relationships between metabolites and samples may happen, the relationship alone is not sufficient to mine the data. To give a complete biological meaning to the data, it is necessary to add biological knowledge on the analyzed metabolites and to their anabolic and catabolic pathways, to understand if the found relationships are due to the metabolic pathways cascade, to indirect causes or are completely random. External databases, like BioCyc (http://biocyc.org) and Kegg (http://www.genome.jp/kegg) can be used for this purpose. Their use can be coupled with statistical software and packages to improve pathway analysis.

Metabolomics data can be often coupled with other types of data, like Trascriptomics, Genomics and Proteomics, to give an insightful meaning to the metabolites levels. The combined use of multiple dataset is called Data Fusion (Smilde et al. 2005). It is based on multivariate statistical analysis of different datasets, and infers on the inner relationships between the datasets.

Deeper description of the data mining process is demanded to literature (Banimustafa & Hardy, 2012).

4.7 Data sharing

In metabolomics, sharing of the data is a key step to strengthen the results found in the laboratory experiment. The validation of the findings of the metabolomics experiments passes through the test of multiple experimental conditions, different kind of analysis, huge amount of observation and variables, different data analysis approaches and data interpretations. Not all these steps can be performed in single studies, which often cover only partially the question under analysis. Furthermore, data analysis approaches are quickly evolving and the same data produced now, might be used for better and more insightful data analysis in the future.

Back in 2007, the Metabolomics Standards Initiative (MSI, Sansone et al. 2007, Fiehn et al. 2007) came out with a call for standardization of the metabolomics studies under precise guidelines for the metabolomics community. One of the suggested procedures was to share the raw data and/or metadata of the experiments with the external reader/researchers, to aloud external researchers to compare, contrast and make inferences from the results they obtain in their experiments (Goodacre et al. 2004). For example, I can report my own experience: I had access to the grape metabolome experimental data, acquired in our lab (Mattivi et al. unpublished data); the design of all the experiments described in this thesis have been all inspired from this data.

The European Bioinformatics Institute (EBI) recently established a database for metabolomics data (Haug et al. 2013). It is called "Metabolights" (http://www.ebi.ac.uk/metabolights/), and it is based on data sharing of the metabolomics data under certain standard procedures. Before storing the data, the analysis must be described in deep using a Tab-based software (ISA-Tab); in this file, all the samples, the methods, and the instruments are described. Raw data and metadata can be stored in the database, together with their Tab file describing how this data has been generated.

The results of the work described in chapter 7 will be stored in Metabolights database, as soon as it will be accepted as manuscript. In our laboratory, the storage of the data is already an automatic process as described by Franceschi et al. (2014). Indeed, information about the data in our lab is stored through ISA-Tab files and after data acquiring, the data is processed through an automatic pipeline specifically designed to work smoothly with Metabolights.

References Chapter 4

1. *Arapitsas, P., Scholz, M., Vrhovsek, U., Di Blasi, S., Biondi Bartolini, A., Masuero, D., … Mattivi, F. (2012). A metabolomic approach to the study of wine micro-oxygenation. PloS One, 7(5), e37783. doi:10.1371/journal.pone.0037783*
2. *Arapitsas, P., Speri, G., Angeli, A., Perenzoni, D., & Mattivi, F. (2013). The influence of storage on the "chemical age" of red wines. Metabolomics, 10(5), 816–832. doi:10.1007/s11306-014-0638-x*
3. *BaniMustafa A. H., Hardy N. W. (2012). A Strategy for Selecting Data Mining Techniques in Metabolomics. Methods in Molecular Biology, 860, pp 317-333*
4. *Benjamini Y, Hochberg Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B Methodol. Jan 1;57(1):289–300.*
5. *Boswell, P. G., Schellenberg, J. R., Carr, P. W., Cohen, J. D., & Hegeman, A. D. (2011). Easy and accurate high-performance liquid chromatography retention prediction with different gradients, flow rates, and instruments by back-calculation of gradient and flow rate profiles. Journal of Chromatography. A, 1218(38), 6742–9. doi:10.1016/j.chroma.2011.07.070*
6. *Bylesio, M., Rantalainen, M., Cloarec, O., Nicholson, J. K., Holmes, E., & Trygg, J. (2007). OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. Journal of Chemometrics, (February), 398–405. doi:10.1002/cem*
7. *Conde, C., Silva, P., Fontes, N., Dias, A. C. P., Tavares, R. M., Sousa, M. J., … Gerós, H. (2007). Biochemical changes throughout grape berry development and fruit and wine quality. Food, 1, 1–22. Retrieved from http://hdl.handle.net/1822/6820*
8. *Coombe, B. G., & McCarthy, M. G. (2000). Dynamics of grape berry growth and physiology of ripening. Australian Journal of Grape and Wine Research, 6, 131–135. doi:10.1111/j.1755-0238.2000.tb00171.x*
9. *Creek, D. J., Jankevics, A., Breitling, R., Watson, D. G., Barrett, M. P., & Burgess, K. E. V. (2011). Identification by Retention Time Prediction, 8703–8710.*
10. *De Livera, A. M., Dias, D. A., Souza, D. De, Rupasinghe, T., Tull, D. L., Roessner, U., … Speed, T. P. (2012). Normalising and integrating metabolomics data Normalising and integrating metabolomics data. Anal. Chem.*
11. *Everitt, B.S. (1993) Cluster Analysis. Edward Arnold, London.*

12. *Feise, R. J. (2002). Do multiple outcome measures require p-value adjustment? BMC Medical Research Methodology, 2, 8. doi:10.1186/1471-2288-2-8*
13. *Fiehn, O., Robertson, D., Griffin, J., vab der Werf, M., Nikolau, B., Morrison, N., … Sansone, S. A. (2007). The metabolomics standards initiative (MSI). Metabolomics, 3, 175–178. doi:10.1007/s11306-007-0070-6*
14. *Franceschi, P., Mylonas, R., Shahaf, N., Scholz, M., Arapitsas, P., Masuero, D., … Wehrens, R. (2014). MetaDB a Data Processing Workflow in Untargeted MS-Based Metabolomics Experiments. Frontiers in Bioengineering and Biotechnology, 2(December), 72. doi:10.3389/fbioe.2014.00072*
15. *Frederiksen, R. B. (2011). Optimize Peak Detection & Integration with ApexTrack / Processing Theory. Waters, 1–59.*
16. *Goodacre, R., Vaidyanathan, S., Dunn, W. B., Harrigan, G. G., & Kell, D. B. (2004). Metabolomics by numbers: Acquiring and understanding global metabolite data. Trends in Biotechnology, 22(5), 245–252. doi:10.1016/j.tibtech.2004.03.007*
17. *Haug, K., Salek, R. M., Conesa, P., Hastings, J., De Matos, P., Rijnbeek, M., … Steinbeck, C. (2013). MetaboLights - An open-access general-purpose repository for metabolomics studies and associated meta-data. Nucleic Acids Research, 41(October 2012), 781–786. doi:10.1093/nar/gks1004*
18. *Helen G Gika, Georgios A Theodoridis, Mark Earll & Ian D Wilson (2012). A QC approach to the determination of day-to-day reproducibility and robustness of LC–MS methods for global metabolite profiling in metabonomics/metabolomics. Bioanalysis, Vol. 4, No. 18, Pages 2239-2247 , DOI 10.4155/bio.12.212*
19. *Jolliffe, I.T. (2002). Principal Component Analysis. Springer series in statistics, ISBN 978-0-387-22440-4*
20. *Kind, T., & Fiehn, O. (2006). Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. BMC Bioinformatics, 7, 234. doi:10.1186/1471-2105-7-234*
21. *Kind, T., & Fiehn, O. (2007). Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. BMC Bioinformatics, 8, 105. doi:10.1186/1471-2105-8-105*
22. *Knolhoff, A. M., Callahan, J. H., & Croley, T. R. (2014). Mass accuracy and isotopic abundance measurements for HR-MS instrumentation: capabilities for non-targeted*

analyses. Journal of the American Society for Mass Spectrometry, 25(7), 1285–94. doi:10.1007/s13361-014-0880-5

23. *Kuhl, C., Tautenhahn, R., Bo, C., Larson, T. R., & Neumann, S. (2012). CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets. Analytical Chemistry, 84, 283–289. doi:10.1021/ac202450gMcGeachie, M. J., Chang, H. H., & Weiss, S. T. (2014). CGBayesNets: Conditional Gaussian Bayesian Network Learning and Inference with Mixed Discrete and Continuous Data. PLoS Computational Biology, 10(6). doi:10.1371/journal.pcbi.1003676*
24. *Patty, G. J., Yanes, O., & Siuzdak, G. (2013). Metabolomics: The apogee of the omics trilogy. International Journal of Pharmacy and Pharmaceutical Sciences, 5(4), 45–48. doi:10.1038/nrm3314*
25. *Robinson, S., Davies, C., & Simon P. Robinson, C. D. (2000). Molecular biology of grape berry ripening. Australian Journal of Grape and Wine Research, 6(Coombe), 175–188. doi:10.1111/j.1755-0238.2000.tb00177.x*
26. *Rosipal, R. (2011). Nonlinear partial least squares: An overview. Chemoinformatics and Advanced Machine Learning Perspectives: Complex Computational Methods and Collaborative Techniques, 169–189. Retrieved from http://aiolos.um.savba.sk/~roman/Papers/npls_book11.pdf\npapers2://publication/uuid/55966E8F-2346-4326-B57E-3A1F47835189*
27. *Sansone, S.-A., Fan, T., Goodacree, R., Griffin, J., Hardy, N. W., Daouk, R.-K., … Fiehn. (2007). The metabolomics standards initiative (MSI). Metabolomics, 3(8), 175–178. doi:10.1007/s11306-007-0070-6*
28. *Shahaf, N., Franceschi, P., Arapitsas, P., Rogachev, I., Vrhovsek, U., & Wehrens, R. (2013). Constructing a mass measurement error surface to improve automatic annotations in liquid chromatography/mass spectrometry based metabolomics. Rapid Communications in Mass Spectrometry, 27(21), 2425–2431. doi:10.1002/rcm.6705*
29. *Smilde, A. K., van der Werf, M. J., Bijlsma, S., van der Werff-van der Vat, B. J. C., & Jellema, R. H. (2005). Fusion of mass spectrometry-based metabolomics data. Analytical Chemistry, 77(20), 6729–36. doi:10.1021/ac051080y*
30. *Smith, C. a, Want, E. J., O'Maille, G., Abagyan, R., & Siuzdak, G. (2006). XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak*

alignment, matching, and identification. Analytical Chemistry, 78(3), 779–87. doi:10.1021/ac051437y

31. *Stanstrup, J., Gerlich, M., Dragsted, L. O., & Neumann, S. (2013). Metabolite profiling and beyond: approaches for the rapid processing and annotation of human blood serum mass spectrometry data. Analytical and Bioanalytical Chemistry, 405(15), 5037–48. doi:10.1007/s00216-013-6954-6*
32. *Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. a., … Viant, M. R. (2007). Proposed minimum reporting standards for chemical analysis. Metabolomics, 3(3), 211–221.*
33. *Suvitaival, T., Rogers, S., & Kaski, S. (2014). Stronger findings for metabolomics through Bayesian modeling of multiple peaks and compound correlations. Bioinformatics, 30, i461–i467. doi:10.1093/bioinformatics/btu455*
34. *Sysi-Aho, M., Katajamaa, M., Yetukuri, L., & Oresic, M. (2007). Normalization method for metabolomics data using optimal selection of multiple internal standards. BMC Bioinformatics, 8, 93. doi:10.1186/1471-2105-8-93*
35. *Theodoridis, G., Gika, H., Franceschi, P., Caputi, L., Arapitsas, P., Scholz, M., … Mattivi, F. (2011). LC-MS based global metabolite profiling of grapes: solvent extraction protocol optimisation. Metabolomics, 8(2), 175–185. doi:10.1007/s11306-011-0298-z*
36. *Thurman, E. M., & Ferrer, I. (2010). The isotopic mass defect: A tool for limiting molecular formulas by accurate mass. Analytical and Bioanalytical Chemistry, 397(7), 2807–2816. doi:10.1007/s00216-010-3562-6*
37. *Vinaixa, M., Samino, S., Saez, I., Duran, J., Guinovart, J. J., & Yanes, O. (2012). A Guideline to Univariate Statistical Analysis for LC/MS-Based Untargeted Metabolomics-Derived Data. Metabolites, 2(4), 775–795. doi:10.3390/metabo2040775*
38. *Warwick B Dunn, Ian D Wilson, Andrew W Nicholls & David Broadhurst (2012). The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans. Bioanalysis, Vol. 4, No. 18, Pages 2249-2264 , DOI 10.4155/bio.12.204*
39. *Wilcoxon, F. (1946). Individual comparisons of grouped data by ranking methods. Journal of Economic Entomology, 39(6), 269. doi:10.2307/3001968*
40. *Wold H. (1974). Causal flows with latent variables ☆: Partings of the ways in the light of NIPALS modelling. European economic review, 5, 67-86.*

41. *Wold, S., Kettaneh-Wold, N., & Skagerberg, B. (1989). Nonlinear PLS modeling. Chemometrics and Intelligent Laboratory Systems, 7, 53–65. doi:10.1016/0169-7439(89)80111-X*
42. *Wold, S., Sjostrom, M., & Eriksson, L. (2001). PLS-regression : a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems, 58, 109–130.*

5. Fusion of GC/MS and LC/HRMS data to improve the identification and confirmation of the unknown Volatile-Aroma-compound precursors' in Grape

Volatile-aroma-compound precursors (or volatile precursors) are a very important class of chemical compounds that determine the quality of the grapes, and their potential in wine production. Despite their importance, their presence in the spectra of the untargeted Reverse Phase LC-MS methods is often not taken in consideration. In this work, using the untargeted LC-MS method developed in our lab (Arapitsas et al. 2013), I demonstrated that many peaks present in the spectra can be identified as volatile precursors at MSI level 2, and that their presence can be at least semi-quantified.

This project has been developed in collaboration with Manoj Ghaste and it is reported in the paper: "Ghaste M.; Narduzzi L.; Carlin S.; Vrhovsek U.; Shulaev V.; Mattivi F, (2015) Chemical Composition of Volatile Aroma Metabolites and their Glycosylated Precursors Uniquely Differentiates Individual Grape Cultivars" (in press). The part of the project here described has been carried out completely by **me**; in the project, I use as reference the data from the GC-MS analysis produced by Manoj Ghaste. Dr. Fulvio Mattivi supervised this project.

5.1 Introduction

Volatile-aroma-compounds are a set of multiple classes of organic compounds that contribute together to the formation of the aroma of the grape berry. Every grape is very rich in Volatile-aroma-compounds, and the ratio between the concentrations of these compounds determines the specific aroma of the different grape varieties. These compounds may play very different roles in plants, like for example as attractant for insects/mammals. They are also reported to be a signaling answer to biotic and abiotic stresses and to be involved in inter-plant signaling (Lund & Bohlmann 2006, Baldwin et al. 2006).

In general, plants tend to accumulate such compounds also in "bound" forms, to store them in the vacuole, ready to be released whenever they would be necessary. In facts, in their "free" form, they cannot be long-term stored, because of their high volatility. The most common way to store these compounds is to bind the "free" form to glycosides. One of the main important roles of the glycosides is the storage of "secondary" metabolites (within there are volatiles too). Indeed, Glycosides increase

the water solubility of the "bound" compounds, binding through a hydrolytic bond to the "free" volatiles (generally called Glyco-conjugated volatiles GCVs). The hydrolytic bond is generally considered weak, because it requires low kinetic energy to be broken, and the reaction carries out without modifying the chemical structure of the released compounds. In grape, the concentration of the volatiles in "bound" glycosidic forms has been reported to be between 2-8 higher than the concentration of their "free" counterparts (Wang et al. 2008, Sarry & Gunata 2004). This means that grape has a very high aroma potential stored in the vacuole of the cells in the berry, especially in the skin, and in minor extent, in the flesh (Robinson et al. 2013).

In winemaking, the reservoir of aroma precursors is one of the most important parameter for the production of quality wines. Indeed, if the free forms get mostly lost during the various step of the vinification, their bound counterpart supply to their loss, continuously releasing new free volatiles. The hydrolysis of the bound forms is not a random process, but it is due in minor extent to the mild-acidic conditions of the vinification process, and overall driven by the hydrolytic enzymes produced by the yeast and bacteria, during fermentation. This means that the final aroma of a specific wine is a combination of the reservoir of the volatiles precursors and the ability of the microbes to hydrolyze the bonds of the glycosidic forms. In some extent, starting from the same grape material, different yeasts may give different aroma, especially if they are from different species (Ciani et al. 2010). Therefore, to study the aroma potential of a grape variety, we need to know the composition of the **intact** "bound" fraction in grape, and their hydrolysis from the various yeast strains.

Many studies have been performed on the bound fraction of the volatiles precursors in grapes but not in in their intact form; indeed researchers were mostly using an indirect method: bound fraction hydrolysis followed by GC-MS analysis to study the released free fraction (Maicas et al., 2005; Esti & Tamborra, 2006). The glycosidic part has also been studied using GC-MS, through derivatization. These methods have been the state of the art for many years, but they have two main limits: it is impossible to study many conjugated volatiles together, and, being an indirect method, the formation of artifacts is common (Little et al. 1999, Esti & Tamborra, 2006).

Recently, two direct methods have been developed to determine the amount of conjugated volatiles (Schievano et al. 2013, Flamini et al. 2014), using LC-MS used in combination respectively with NMR and GC-MS. The two methods reported were very efficient; both were able to identify more than ten precursors in their intact forms. On the other hand, the first method had a very long process of sample preparation, and the isolation necessary for the NMR analysis of the compounds is a time consuming step, which has not been considered in my work. The second method is conceptually similar

to the work reported here. However, differently from this work, the authors focused only on a chemical class of compounds (terpenoids), developing specific extraction and chromatographic methods to achieve their goals, excluding from the analysis all the other volatile compounds that might be accumulated in the grapes.

The aim of this work was to use the information about the hydrolyzed volatiles contained in the analyzed grapes from the GC-MS analysis (after hydrolysis), to putatively identify their glyco-conjugated precursors through LC-HRMS and MS/MS analyses and confirm the identifications through intensities correlation between the two analytical platforms. This method can be considered as an untargeted analysis to identify putatively unreported volatile precursors in grape.

5.2 Materials and Methods

5.2.1 Grape samples:

The grapes used in this experiment are shown in table 1. All the samples have been collected in the experimental fields of the "Fondazione Edmund Mach" in San Michele all'Adige (TN) Italy, during the season 2013. They were harvested at technical maturity, 18° brix, immediately frozen under liquid nitrogen and stored at -80° until analysis. 1 Kg of healthy grape were powdered in liquid nitrogen using an analytical mill (IKA® -Werke GMbH & Co. Staufen, Germany) prior to sample preparation. Every sample has prepared in triplicate.

No	Prime name	Berry Color	Origin	Short name
1	Riesling	White	Vinifera	RIE
2	Gewürztraminer	Pink	Vinifera	GWT
3	Moscato rosa	Red	Vinifera	MOR
4	Iasma ECO 1	White	Vinifera	F3P30
5	Iasma ECO 2	White	Vinifera	F3P63
6	Iasma ECO 3	White	Vinifera	F3P51
7	Nero	Red	Hybrid variety	NERO
8	Isabella	Red	Hybrid variety	ISA
9	*Vitis arizonica*	Red	American	VAT
10	*Vitis cinerea*	Red	American	VCI

Table 1: The grape material used in this experiment. All the material is from the experimental field of the Fondazione Edmund Mach in San Michele all'Adige, (TN) Italy.

5.2.2 Chemical reagents:

Methanol, dichloromethane, formic acid and pentane were purchased from Sigma Aldrich (Milan, Italy). Sodium sulphate anhydrous and citric acid were purchased from Carlo Elba (Milan, Italy). The water used in the experiments was purified with a Milli-Q water purification system from Millipore (Bedford, MA, USA), SPE cartridges Isolute ENV+ (1g, 6mL) were obtained from Biotage (Uppsala, Sweden), Rapidase® AR2000 enzyme was purchased from DSM Food Specialties B.V. (Delft, The Netherlands).

5.2.3 Sample preparation:

The experimental design is summarized in Image 1. The method used is the same of Vhroscek et al. (2014) slightly modified. 30 g of grape powder, 80 mL water and 0.5 g of gluconolactone were taken and 25 µL of 1-Heptanol was added as internal standard (1257 mg/L). The solution was then homogenized for 3 min at 20000 rpm using an ultra-Turrax homogenizer, followed by centrifugation for 5 min. at 10000 rpm at 5 °C. The supernatant obtained was then filtered through filter paper and the

extract was further used for the SPE procedure. Isolute ENV+ cartridges were conditioned with 20 mL each of methanol and equilibrated with 20 ml of milliQ water, then the grape extract was loaded and eluted through cartridges and the cartridges were washed with 20 mL of water to remove water soluble impurities. Free volatiles were eluted with 20 mL of dichloromethane, the elute was collected in a glass tube and 40 mL of pentane was added to it. This solution was dried with anhydrous Na2SO4 and concentrated to 200µL as described in (Boido et al., 2003) and successively injected in the GC-MS. The conjugated volatile precursors were eluted with 30ml of Methanol and 1mL of methanol was diluted with 1 ml of water and injected in the LC-MS system for the precursor's analysis; the rest of the fraction was evaporated to dryness. Then the flask containing it was rinsed with 10 mL of dichloromethane to remove any leftover free volatile compounds. The bound fraction was then re-dissolved in 5 mL of citrate buffer at pH 5 and 200µL of enzyme AR2000 (70 mg/mL, supply by: "Oenobrands") was added and kept in a 40°C water bath for 48 Hrs. Later, 10µL of internal standard 1-Heptanol was added; 1 ml of this fraction was diluted with 3 ml of Methanol and injected in the LC-MS system while with the remaining fractions, free volatiles were extracted with 3mL of pentane/dichloromethane 2:1, v/v, three times, all organic phase containing hydrolyzed volatiles being concentrated carefully to a volume of 200µL for GC-MS analysis. All samples were prepared and measured in three replicates.

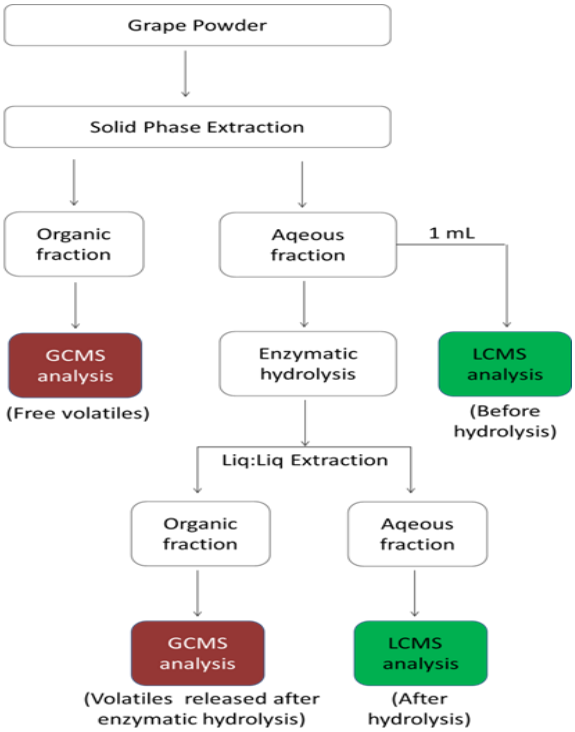


Image 1: the diagram displays the experimental design. The basic principle behind this experiment is that The LC-MS analysis (before hydrolysis) is used to individuate the glycosidic precursors of the volatiles measured by the GC-MS analysis (after hydrolysis). Then, in the LC-MS analysis (after hydrolysis) the compound identified as putative precursors, should disappear or at least decrease in concentration in a manner relative to the amount measured in the GC-MS instrument.

5.2.4 LC-MS analysis and data extraction

LC-HRMS analysis was performed using the "UPLC-Synapt" instrument (UHPLC-ESI-Q-IMS-TOF-MS) described in the section 2.3.2 of this thesis. 1 ml of Methanolic elute (before hydrolysis) was diluted with 1 ml of water and filtered through 0.22 um Millisart filters (Whattman, Milano, Italy) and injected in partial loop needle overfill mode. 0.5 mL of aqueous samples (after hydrolysis) was diluted with 1.5 ml of methanol filtered and injected as for the previous samples. All LC and MS instrumental settings were kept the same as described in the chapter 2 of this thesis. Raw data was extracted using the MetaDB and MetaMS pipeline described by Franceschi et al. (2014) based on XCMS software coupled to the CAMERA package (Smith et al., 2006, Kuhl et al. 2012). The data was normalized using the Total ion current normalization (Alfassi et al. 2004).

5.2.5 MS/MS analysis

One sample per each variety was re-injected under the same chromatographic condition in the UPLC-Synapt instrument. LC-MS/MS analysis was performed selecting the precursor ions in the quadrupole and fragmenting them in the transfer sector, as described in chapter 2. Fragmentation was obtained using a collision energy profile from 20 to 25 eV; low collisional energy has been adopted because the hydrolytic bond is a weak bond and glycoside cleavage is obtained at low CID. Acquired MS/MS spectra were manually integrated and queried in MetFrag (Wolf et al. 2010) against KEGG (www.kegg.jp), PubChem(https://pubchem.ncbi.nlm.nih.gov/) or Chemspider (www.chemspider.com) databases, or using the GCV structure database described in section 5.3.1.

5.2.6 Pearson correlation analysis

Pearson correlation analysis has been performed between the LC-MS putative identified peaks, and their corresponding released volatile forms from the GC-MS analysis (Flamini et al. 2014). The area of the putative identified peaks measured after hydrolysis was subtracted from the area of the same peak before hydrolysis, obtaining the hydrolyzed amount of such peak. The obtained values, have been correlated to their corresponding volatiles previously measured in GC-MS.

5.3 Results and discussion

5.3.1 Filtering the data

The list of volatiles released (after hydrolysis) from GC-MS analysis (Ghaste et al. 2015 and table 2) was used to build all the possible structures of the glyco-conjugated volatiles (GCV structures database) using the literature (Fernandez-Gonzalez & Di Stefano 2004, Sarry & Gunata 2004) to select the proper glycosides to bond to the volatiles and build an in-silico library of GCV structures. For all the GCV structures, the mono-isotopic masses (MM) were calculated, here reported in table 2. From the MM, all the m/z adducts in positive mode ($+H^+$, $+Na^+$, $+K^+$), and in negative mode ($-H^-$, $+Cl^-$, $+FA^-$) were calculated. This GCVs m/z list has been used to select XCMS peaktable features as described below.

The selection of the peaktable features was obtained applying multiple filters here listed: 1) RT filter, 2) m/z range, 3) only odd m/z ions (Nitrogen rule) 4) relative mass defect range (RMD), 5) intensity threshold, 6) de-isotoping.

1) The GCV structures are composed of an hydrophobic part (the volatile) and an hydrophilic one (the glycosides); so their logD cannot be extremely different and most likely they elute only inside a given RT range (relationship between logD and RT is explained in chapter 2). Comparing their calculated logD with the ones of known metabolites, was possible to establish a broad range RT filter. LogD have been calculated using the "ACD/Labs 12.0 Percepta Platform - PhysChem Module" setting as pH 2.87 (measured pH for 0.1% formic acid in water); Benzyl-primeveroside showed the lowest calculated logD value: -0.17 which was similar to the one calculated for Epicatechin (-0.10), that has a RT of 12.7 minutes in our chromatographic method. So, considering the error, the lowest limit of the RT filter was settled at 10 minutes. Geranyl-arabinopyranosyl-glucoside was showing the highest logD value, 1.27, similar to the one calculated for quercetin-glucoside (1.3) that has a RT of 19.8 minutes. So the highest limit was settled at 23 minutes.
2) m/z range was simply calculated from the GCV m/z list, taking as limits the lowest and the highest possible m/z. The selected range was from 250 to 600 Dalton.
3) The features with an even m/z mass have been excluded, because all the compounds did not contain Nitrogen in their formula, so, according to the Nitrogen rule, their MM cannot be odd (and their ion cannot have an even m/z).

4) RMD range was simply calculated from the GCV m/z list as well, taking as limits the lowest and the highest possible RMD. The lowest bond was 300 ppm, while the highest bond was 600 ppm. Theory about the RMD is explained in chapter 6 and in literature (Sleno, 2012).
5) An intensity threshold is necessary to perform MS/MS analysis. Ions below 1000 counts/sec are not suitable for fragmentation in the "Synapt" and were discarded.
6) Ions being part of the same isotopic pattern of other ions have been excluded, keeping only the mono-isotopical ions in the final list. The "Isotopes" have been recognized using the isotope list produced by CAMERA, and a further manual checking.

In the XCMS peaktable there were around ≈13 thousands features. Among these, only 403 features matched the filtering requirements described above. The rest of the features have been excluded. I took in account that the exclusion of some data based on arbitrary filters may lead to false negative exclusions. On the other hand, the selection of a limited number of features allows performing a user-dependent MS/MS analysis that can enforce or exclude putative GCV features.

A different selection of the putative features has been described in literature (Tikunov et al. 2011); in that work, the authors selected the features fusing the data from the GC-MS on aroma released from the tomatoes and the features measured in LC-MS. After fusion they performed a PCA and they selected for MS/MS analysis only the features displaying a small Euclidean distance from the Aroma compounds measured from the GC-MS. This choice allowed the authors to identify multiple precursors of tomato volatile aroma compounds and to identify the different kinds of possible glycosylation.

Nevertheless, in my work their approach was not possible. They had a more favorable samples to features ratio (94/1000) in comparison to mine (30/13000) and in their experiment the hydrolysis was complete, while in my experiment the hydrolysis was the bottleneck of the experiment, and will be discussed deeper in the section 5.3.7.

Component Name	MM	glucose	malonyl-glucos	pentosyl-gluco	deoxy-glycosyl-gluco	Glycosyl-gluco
2-Hexenol	100.0888	262.1412	348.1416	394.1840	408.1982	424.1931
trans 3-hexenol	100.0888	262.1412	348.1416	394.1840	408.1982	424.1931
n-hexanol	102.1044	264.1568	350.1572	396.1996	410.2138	426.2087
benzyl alcohol	108.0575	270.1099	356.1103	402.1527	416.1669	432.1618
Hexanoic acid	116.0837	278.1361	364.1365	410.1789	424.1931	440.1880
beta Phenyl ethanol	122.0731	284.1255	370.1259	416.1683	430.1825	446.1774
Anisyl alcohol	124.0524	286.1048	372.1052	418.1476	432.1618	448.1567
Furaneol	128.0473	290.0997	376.1001	422.1425	436.1567	452.1516
ethyl- 3 hydroxy-butanoate	132.0786	294.1310	380.1314	426.1738	440.1880	456.1829
Chavicol	134.0732	296.1256	382.1260	428.1684	442.1826	458.1775
Cinnamyl alcohol	134.0732	296.1256	382.1260	428.1684	442.1826	458.1775
α-Cumyl alcohol	136.0888	298.1412	384.1416	430.1840	444.1982	460.1931
4-Methoxyphenethyl alcohol	152.0837	314.1361	400.1365	446.1789	460.1931	476.1880
4-vinylguaiacol	152.0837	314.1361	400.1365	446.1789	460.1931	476.1880
Hotrienol	152.1201	314.1725	400.1729	446.2153	460.2295	476.2244
4-terpineol	154.1357	316.1881	402.1885	448.2309	462.2451	478.2400
alpha-terpineol	154.1357	316.1881	402.1885	448.2309	462.2451	478.2400
cis Geraniol	154.1357	316.1881	402.1885	448.2309	462.2451	478.2400
linalool	154.1357	316.1881	402.1885	448.2309	462.2451	478.2400
p-Menth-8-en-3-ol	154.1357	316.1881	402.1885	448.2309	462.2451	478.2400
trans Geraniol	154.1357	316.1881	402.1885	448.2309	462.2451	478.2400
β-Citronellol	156.1514	318.2038	404.2042	450.2466	464.2608	480.2557
Ethyl 3-hydroxyhexanoate	160.1095	322.1619	408.1623	454.2047	468.2189	484.2138
eugenol	164.0837	326.1361	412.1365	458.1789	472.1931	488.1880
Acetovanillone	166.0629	328.1153	414.1157	460.1581	474.1723	490.1672
Homovanillyl alcohol	168.0786	330.1310	416.1314	462.1738	476.1880	492.1829
Geranic acid	168.1150	330.1674	416.1678	462.2102	476.2244	492.2193
7 OH geraniol	170.1306	332.1830	418.1834	464.2258	478.2400	494.2349
8 OH linalool cis	170.1306	332.1830	418.1834	464.2258	478.2400	494.2349
8 OH linalool trans	170.1306	332.1830	418.1834	464.2258	478.2400	494.2349
exo-2-Hydroxycineole	170.1306	332.1830	418.1834	464.2258	478.2400	494.2349
lilac alcohol A	170.1306	332.1830	418.1834	464.2258	478.2400	494.2349
lilac alcohol B	170.1306	332.1830	418.1834	464.2258	478.2400	494.2349
lilac alcohol C	170.1306	332.1830	418.1834	464.2258	478.2400	494.2349
Linalool oxide A	170.1306	332.1830	418.1834	464.2258	478.2400	494.2349
Linalool oxide B	170.1306	332.1830	418.1834	464.2258	478.2400	494.2349
Linalool oxide C	170.1306	332.1830	418.1834	464.2258	478.2400	494.2349
Linalool oxide D	170.1306	332.1830	418.1834	464.2258	478.2400	494.2349
OH nerol	170.1306	332.1830	418.1834	464.2258	478.2400	494.2349
p-menth-1-ene-7,8-diol	170.1306	332.1830	418.1834	464.2258	478.2400	494.2349
Terpendiol I	170.1306	332.1830	418.1834	464.2258	478.2400	494.2349
Terpendiol II	170.1306	332.1830	418.1834	464.2258	478.2400	494.2349
hydroxy Citronellol	172.1463	334.1987	420.1991	466.2415	480.2557	496.2506
Coniferol	180.0786	342.1310	428.1314	474.1738	488.1880	504.1829
Homovanillinic acid	182.0579	344.1103	430.1107	476.1531	490.1673	506.1622
3,4,5-Trimethoxyphenol	184.0735	346.1259	432.1263	478.1687	492.1829	508.1778
Scopoletin	192.0423	354.0947	440.0951	486.1375	500.1517	516.1466
3 hydroxy beta damascone	208.1463	370.1987	456.1991	502.2415	516.2557	532.2506
4-oxo beta ionol	208.1463	370.1987	456.1991	502.2415	516.2557	532.2506
3,4-dihydroactinidol	210.1619	372.2143	458.2147	504.2571	518.2713	534.2662
Dihydro-3-oxo-beta-ionol	210.1620	372.2144	458.2148	504.2572	518.2714	534.2663
vomifoliol	224.1412	386.1936	472.1940	518.2364	532.2506	548.2455
3-4-dihydro-3-oxoactinidiol I	226.1563	388.2087	474.2091	520.2515	534.2657	550.2606
3-4-dihydro-3-oxoactinidiol II	226.1564	388.2088	474.2092	520.2516	534.2658	550.2607
3-4-dihydro-3-oxoactinidiol III	226.1565	388.2089	474.2093	520.2517	534.2659	550.2608

Table 2: The list of all the possible GCV monoisotopic masses, obtained coupling the volatiles list from Ghaste et al. (2015) and the conjugated glycoside reported in literature.

5.3.2 Matching

The 403 features have been matched back to the GCV m/z list, to match each m/z to its corresponding putative GCV structure. An error of ± 0.005 Dalton has been considered acceptable. A total amount of 147 features were matching with one or more GCV structures; these features were selected for MS/MS analysis. Within the remaining non-matching features, 20 of them with the highest Intensity value have been selected as well for the MS/MS analysis, for a total of 167 MS/MS analysis. Many of the remaining ions have been manually checked. Most of them were noisy, with an unclear peak shape different sample by sample.

5.3.3 MS/MS analysis

All the MS/MS spectra acquired in the experiments have been firstly checked for correctness. As stated in chapter 2, the main limit of the Q-TOF is that the ion filtering done by the quadrupole is able to select ions only with a mass range of ± 0.5 Dalton (1 Dalton window in total). This means that co-eluting ions with m/z inside this mass range, could be selected and analyzed together with the ions of interest (image 2), contaminating their MS/MS spectrum. For this reason the checking was necessary.

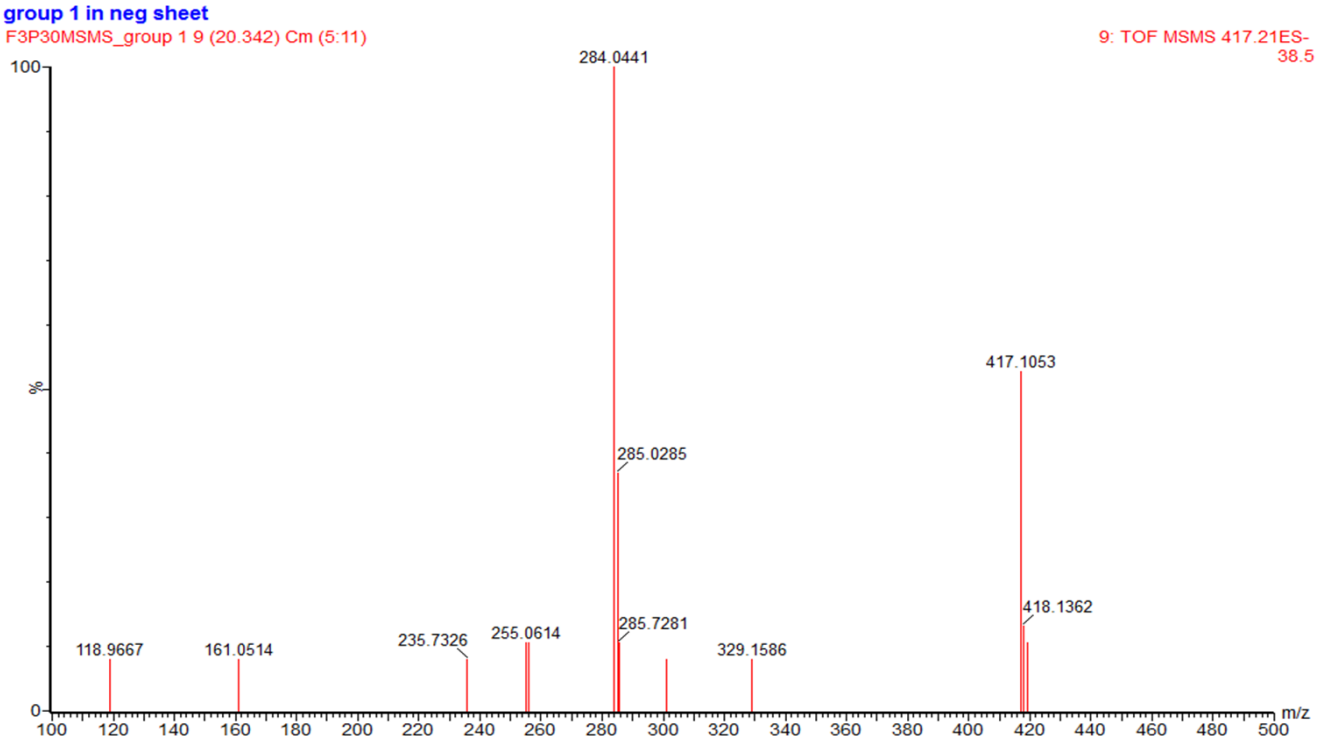


Image 2: An incorrect MS/MS spectrum. The target ion had an m/z of 417.21, but in the chromatogram is clear that the fragments obtained are from both 417.21 and 417.1053 precursor ions. This MS/MS spectrum is not interpretable.

The clean spectra have been queried in MetFrag against KEGG, ChemSpider and PubChem databases. Even if the composition of the volatiles precursors in the plant kingdom is mostly known, their exact structure is still underdetermined in most cases, and these compounds cannot be found largely in the databases. Indeed the query against online databases gave, in many cases, structures very different from the expected ones, mostly because the expected structures are not in such databases. To overcome this problem I queried the MS/MS spectra also against the GCV structures database; whenever a structure from the GCV database was showing clearly a higher number of possible fragments and a higher mass accuracy, the compound was labelled as putative identification. After MS/MS analysis, I obtained 52 positive matching spectra. Every MS/MS spectrum has been acquired from the sample that was showing the highest intensity for the precursor ion in LC-MS analysis.

5.3.4 Post-hydrolysis sample analysis

All the in-silico analysis described previously, have been performed on the samples before the enzymatic hydrolysis. Here I describe how the samples after the enzymatic hydrolysis have been used to understand the hydrolysis efficiency, in two main ways: 1) evaluate the "disappearance" of the signal of hydrolyzed compounds, and 2) explore the possibility of the "appearance" of new peaks in the chromatograms. For appearance, I mean the possibility that the released volatiles could be analyzed directly in LC-MS, using specific techniques to improve their ionization and observe their signals in the chromatograms. Pure chemical volatile standards have been injected in the LC-MS system, to develop the method, and then the same analysis has been performed on the samples after hydrolysis. Results are shown below.

1) The hydrolysis of volatiles precursors showed to be incomplete in many cases, indeed the peaks were not disappearing after hydrolysis, but often only reducing their peak area and showed also to have a distinct efficiency between the different putative precursors. A clear example is shown in the image 3: the three colored peaks showed a similar MS/MS spectra and were identified as putative benzyl-pentosyl-glucosides. After hydrolysis, the peaks at minute 12 and 13.20 were disappearing, while the peak at 12.95 had a smaller peak area, but still was clearly present in the chromatogram. As AR2000 is a mixture of α-Arabinosidase, Apiosidase, Rhamnosidase, and β-glucosidase (www.oenobrands.com), I supposed that the peak at 12.95 is a benzyl-xylosyl-glucoside and that the enzyme had only a side effect on the hydrolysis of such metabolite (a xylosidase would be required for direct effect on the metabolite). Nevertheless, the hydrolysis performed by the Rapidase

AR2000 is only a compromise between different strategies and it is known to be not the golden hydrolyzing technique (that does not exist currently). The implications will be discussed in section 5.3.7.

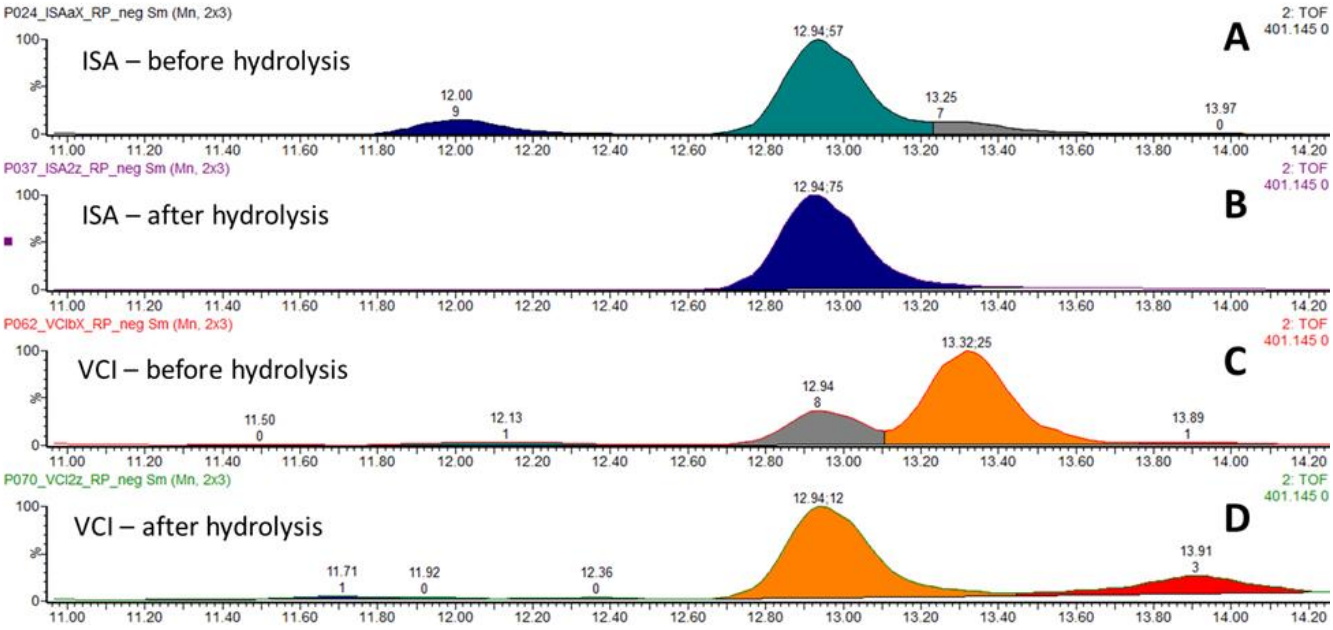


Image 3: A comparison of the EIC of the ion 401.145 before and after hydrolysis in two different samples, VCI (*Vitis* cinerea) and ISA (Isabella). The peak at 12, 12.95 and 13.25 showed similar spectra and were identified all as Benzyl_pentosyl-glucosides. Nevertheless the hydrolysis efficiency was different between the first and third peak versus the second peak.

2) The direct analysis of volatiles in LC-MS is not reported anywhere, because naturally their analysis has been carried out with GC-MS. Personally I was very curious to see if the analysis of free organic volatiles was technically possible with an LC-MS instrument. In theory there should not be barriers in the ionization and detection, since the organic volatiles often resemble structurally smaller versions of bigger analytes already detectable in LC-MS. For example, Hexanoic acid can be considered a short chain version of longer lipidic acids that are currently analyzed in lipidomics analysis (e.g. palmitic acid, oleic acid). The same Benzyl-alcohol can be considered as precursors of more complex metabolites like Benzoic acid, hydroxy-Benzoic acid, Gallic acid and many others, with the only difference that such acids have a scarce volatility. So, the difference in their

detection is imputable to the scarce volatility of the latter, and, as we will see next, in their low concentration in the samples of the former.

As reported by Shahaf et al. (2013), simple Terpenols and their modified versions can be detected in LC-MS chromatograms when their concentration is above 10 mg/L. The first step performed was to observe their possible ionization and the choice of the best ionization method. A Standard mixture of 98 different volatiles (Vhrovsek et al. 2014) have been injected in the LC-MS system using the chromatographic method of Theodoridis et al. (2012), in positive and negative ionization mode (Image 4). APCI ion source has been also tested to analyse these compounds: the idea was that analysing low polarity metabolites, theoretically APCI could improve ions formations of the volatiles as reported in the Image 3 in Chapter 2. So first I tried with Direct Infusion of the volatiles, to observe the ionization performance of the APCI source (Image 5).

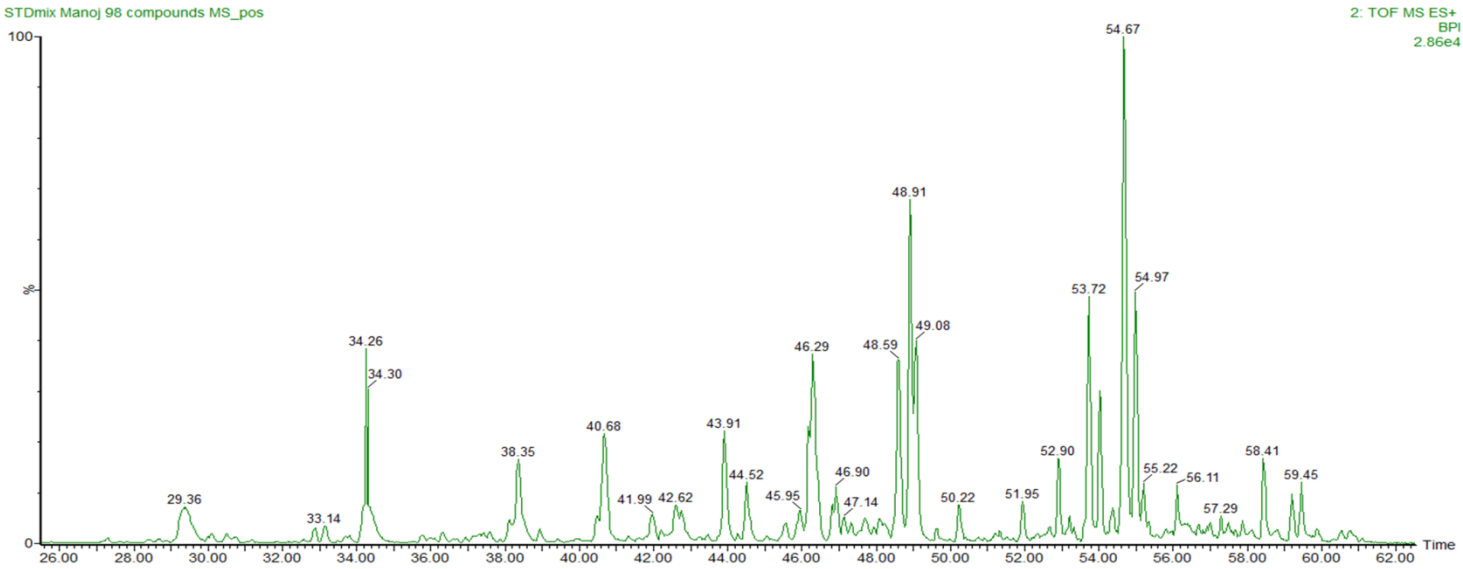


Image 4: Chromatogram of the injection of the volatiles standards mix injected in the UPLC-Synapt system, containing 98 different volatiles commonly analyzed in GC-MS, with concentration ranging from 10 to 100 mg/L. The chromatogram shows ESI+ mode is able to ionize these compounds. The list of the compounds is reported by Vhrovsek et al. (2014). The same STDmix was used to develop a targeted GC-MS method for the detection and quantification of hundreds of grape, strawberry and raspberry metabolites.

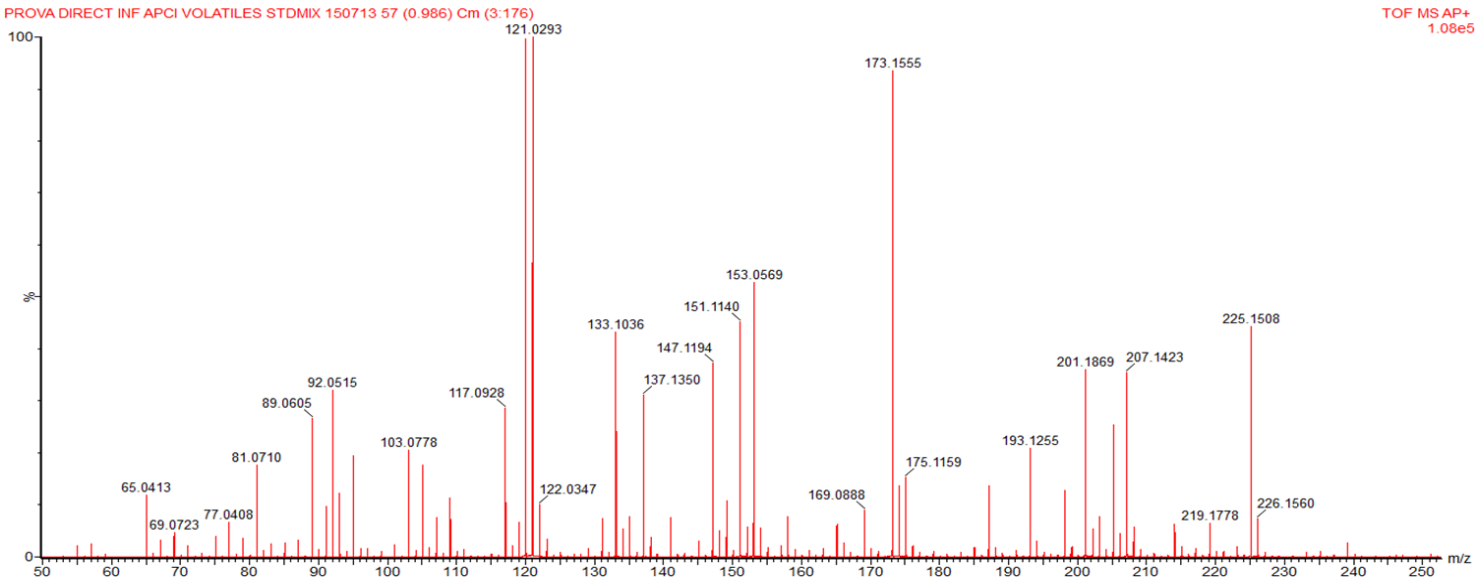


Image 5: The spectrum obtained in the APCI+ source after injecting the volatiles standard mix in direct infusion in the Synapt mass spectrometer. The list of the compounds is reported by Vhrovsek et al. (2014). The same STDmix was used to develop a targeted GC-MS method for the detection and quantification of hundreds of grape, strawberry and raspberry metabolites.

Only positive ion source was able to create clear MS spectra for both ESI and APCI ion sources, while in negative mode, no signal was detected.

The second step was to observe their possible presence in the samples under analysis. The analysis of the samples after hydrolysis with the ESI ion source did not show any clear ion from the volatile compounds. So, I decided to perform the injections also using the APCI source, because it was able to ionize volatile standards and showed a lower signal for polyphenols, which might be suppressing the signal for the volatile compounds in the ESI ion source. APCI was used as source to analyse the samples after hydrolysis using the method of Arapitsas et al. (2014), slightly modifying source settings to adapt it to the APCI. In facts, in the Synapt instrument the interchangeability of the ESI source and APCI source is assured, and most of the settings do not need tuning during source switching. Only the gas flow and gas temperature need to be slightly increased to obtain a slightly better ionization.

The results of this analysis were discouraging; almost none of the STDmix peaks were corresponding to the ones observed in the sample (Image 6); whenever a peak was matching, its intensity was too low to perform a successful MS/MS analysis to confirm the identification. I personally think that this is due to the low amount of free volatiles present in the samples. Indeed, in Ghaste's analysis (Ghaste et al. 2015), the amount of

the compounds was in the range of 10 to 1000 µg/Kg, so around 1000 times less than in the STDmix. Even if I proved that the analysis of volatiles is possible in positive mode with LC-MS instruments both with ESI and APCI source, their direct detection in the samples is far to be achieved, due to sensitivity limits.

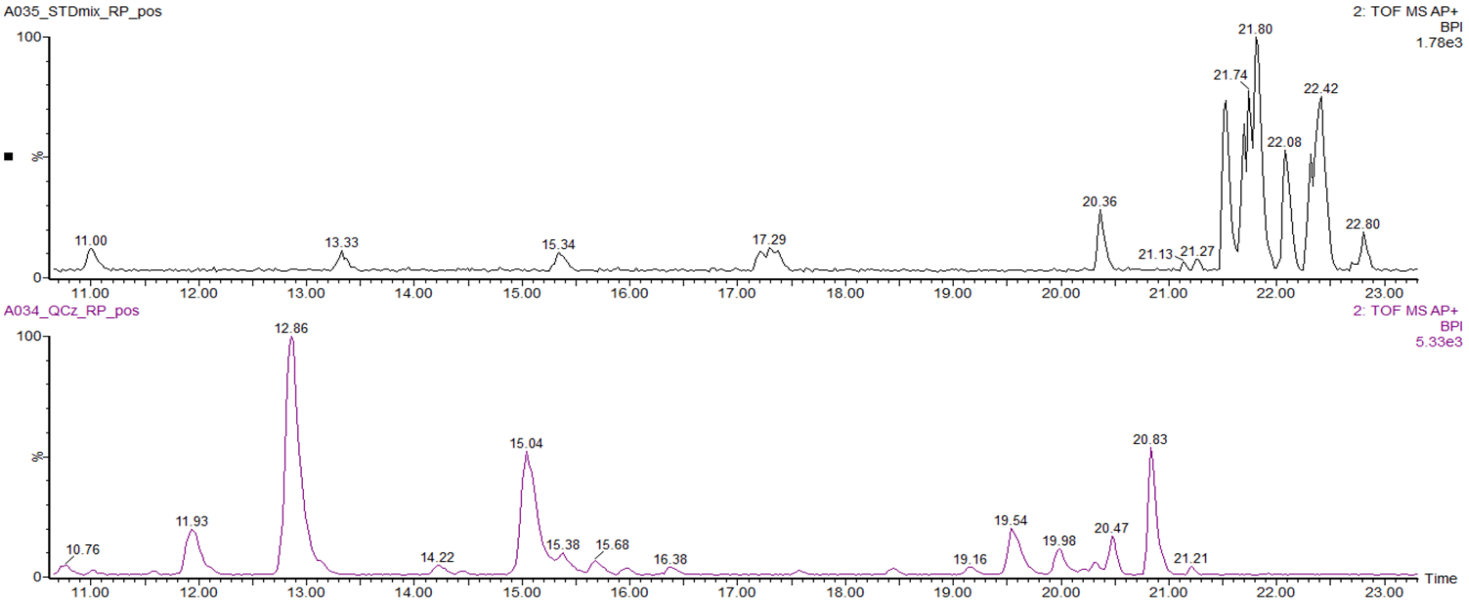


Image 6: A comparison between the chromatograms of the STDmix (up) and the QC (down). The picture shows that despite of the high amount of compounds in the standards mix, none of them could be found in the QC. The figure is only an example. Almost none of the peaks of the sample is exactly matching the RT of the peaks from the STDmix. This was only shown to give a visual impact of the difference between the sample and the STDmix, the two solutions having completely different compositions.

5.3.5 Pearson correlation analysis and peak identification.

The peak area of the 52 putatively identified peaks has been integrated in both samples before and after hydrolysis. In facts, if the peak was present still after the hydrolysis, it meant that was not subject to any clear hydrolysis, and its identification could not be confirmed correlating the area with the GC-MS data and might be wrong. Out of 52 peaks, only 31 corresponded to a unique peak having a clear hydrolysis (at least partial). In most of the cases, the hydrolysis was not complete, and some signal could be observed in the samples after hydrolysis. The reasons why this could happen will be discussed in the section 5.3.7.

To strengthen the identifications, the idea was to correlate the putative identified peaks with their corresponding derivatives, across the whole sample set (30 samples). To overcome the lack of complete hydrolysis, I subtracted the area of the peaks before hydrolysis to the ones after hydrolysis. In total, I

found 15 peaks with a Pearson correlation superior to 0.8. The correlating compounds are displayed in table 3.

N°	Conjugate name	Correlation	Released volatile
1	*ß-phenyl_ethyl_glucosyl-glucoside*		
2	*ß-phenyl_ethyl_arabinosyl-glucoside*	0.91	β-phenyl-ethanol
3	*ß-Phenyl_ethyl_xylosyl-glucoside*		
4	*benzyl_arabinosyl-glucoside*		
5	*benzyl_xylosyl-glucoside*	0.93	Benzyl-alcohol
6	*benzyl_apiosyl-glucoside*		
7	*furaneyl_deoxyhexosyl-glucoside_conjugate*	1.00	Furaneol
8	*trans linalyloxide_hexosyl-glucoside*	0.98	Trans-linalool-oxide
9	*terpenyl_deoxy-hexosyl-glucoside*	0.83	Terpenol
10	*terpenyl_pentosyl-glucoside*		
11	*trans hydroxy-linalyl_deoxy-hexosil-glucoside*	0.97	Trans-hydroxy-linalool
12	*cis hydroxy-linalyl_pentosyl-glucoside*	0.96	Cis-hydroxy-linalool
13	*geranic acid_deoxyhexosyl-glucoside*	0.88	Geranic acid
14	*geranic acid_pentosyl-glucoside*		
15	*vomifolyl_glucoside*	0.88	Vomifoliol

Table 3: The correlating precursors, the R value and the correspondent volatile are displayed. Correaltion has been performed across the whole sampleset. The p-values were always p-value < 0.00001.

During the correlation analysis, the putative precursors corresponding leading to the same volatile have been summed. Also singular correlation have been attempted for such metabolites, but the grouping showed a better results in all cases, expect for the terpenyl_pentosyl-glucoside that had a better correlation with Linalool (0.9) than with the same of all of them (0.826). Nevertheless, the contribution of undetected, unidentified peaks cannot be excluded.

The m/z 485.1859 (-ve mode) showed a unique moiety in its MS/MS spectra, 127.0425, corresponding to Furaneol-H (table 4). Its peak area matched perfectly with the Furaneol concentration in GC-MS (R>0.99). Nevertheless, it was not possible to understand its structure, which shows the typical fragmentation pattern of glycoside moieties, but does not correspond to any previously found in the *Vitis*. The highest concentrations of terpendiols in GC-MS analysis were represented by cis-hydroxy linalool and trans-hydroxy linalool; both were found to correlate strongly with two conjugates identified as cis Hydroxyl-linalool_Pentosyl-Glucoside (R=0.96) and trans-hydroxyl-Linalyl_Rhamnosyl-Glucoside (R=0.97) respectively. A third putative terpendiol conjugate was strongly correlated with trans-linalool oxide (R=0.97) and was putatively assigned as trans-Linalyloxide_Glucosyl-Glucoside. Two peaks clearly showed MS/MS spectra corresponding to 2 different geranic acid conjugates, and the sum of their area also correlated with the concentration

detected using GC-MS (R=0.879). Nevertheless, the hypothesis that other undetected minor peaks may contribute to the accumulation of this compound cannot be discarded.

	NAME	RT	PARENT ION	MS/MS	Formula	Fragment	PARENT ION	MS/MS	formula	Fragment	PARENT ION	MS/MS	formula
1	**Benzyl-alcohol_apiosyl-glucoside**	12.10	401.1453 ([M-H]-)		C18H25O10		447.1503 ([M+FA]-)		C19H27O12				
				269.11	C13H17O6	[M-H-Apiose]		401.14	C18H25O10	[M-FA-H]			
				161.06	C6H9O5	[Glucose – H2O-H]		269.11	C13H17O6	[M-H-FA-Apiose]			
				159.03	C10H7O2			161.05	C6H9O5	[Glucose-H2O-H]			
				143.04	C6H6O4			159.04	C10H7O2				
				131.03	C5H7O4	[Apiose-H2O-H]		131.03	C5H7O4	[Apiose-H2O-H]			
				128.00	C5H404			113.03	C5H5O3				
				113.03	C5H5O3								
				101.03	C4H5O3								
2	**Benzyl-alcohol_xylosil-glucoside**	12.90	401.1448 ([M-H]-)		C18H25O10		447.1500 ([M+FA]-)		C19H27O12				
				269.11	C13H17O6	[M-H-Xylosyl]		401.14	C18H25O10	[M-FA-H]			
				161.06	C6H9O5	Glucose-H2O-H]		269.11	C13H17O6	[M-H-FA-xylose]			
				159.03	C10H7O2			161.06	C6H9O5	[Glucose-H2O-H]			
				143.04	C6H6O4			159.03	C10H7O2				
				131.03	C5H7O4	[Xylose-H2O-H]		131.03	C5H7O4	[Xylose-H2O-H]			
				113.03	C5H5O3			113.03	C5H5O3				
				101.03	C4H5O3			101.03	C4H5O3				
3	**Benzyl-alcohol_arabinosyl-glucoside**	13.35	401.1443 ([M-H]-)		C18H25O10								
				269.11	C13H17O6	[M-H-Arabinosyl]							
				161.06	C6H9O5	[Glucose-H20-H]							
				159.03	C10H7O2								
				131.03	C5H7O4	[Arabinose-H2O-H]							
				113.03	C5H5O3								
				101.03	C4H5O3								
4	**β-phenyl-ethanol-glucosyl-glucoside**	13.04	445.1710 ([M-H]-)		C20H29O11								
				243.05	C10H11O7	[M-H-C10H18O4]							
				162.07	C10H10O2	[M-H-phenyl-ethanol-CH4O]							

				161.05	C6H9O5	**[**Glucose-H2O-H**]**							
				149.05	C5H9O5								
				143.04	C6H7O4								
				119.05	C8H7O1								
				101.02	C4H5O3								
5	**β-phenyl-ethanol-arabinosyl-glucoside**	15.40	415.1606 (**[**M-H**]**-)		C19H27O10		461.165 (**[**M+FA**]**-)		C20H29O12				
				191.06	C7H11O6	**[**Arabinose+C2H4O**]**		415.16	C19H27O10	**[**M-H-FA**]**			
				179.07	C10H11O3	**[**Phenyl ethanol+C2H4O**]**		149.03	C5H9O5	**[**Pentose-H**]**			
				161.05	C6H9O5	**[**Glucose-H2O-H**]**		131.04	C5H7O4	**[**Pentose-H-H2O**]**			
				151.04	C8H7O3			119.04	C8H7O1				
				149.05	C5H9O5	**[**Arabinose-H**]**		101.02	C4H5O3				
				132.04	C5H8O4	**[**Arabinose-H-H2O (A+1)**]**							
				131.04	C5H7O4	**[**Arabinose-H2O-H**]**							
				119.05	C8H7O1	**[**Phenyl-ethanol-2H-H**]**							
				113.02	C5H5O3								
				103.04	C4H7O3								
				102.03	C4H6O3								
				101.02	C4H5O3								
6	**β-phenyl-ethanol-xylosyl-glucoside**	15.90	415.1606 (**[**M-H**]**-)		C19H27O10								
				191.06	C7H11O6	**[**Xylose+C2H4O**]**							
				179.07	C10H11O3	**[**Phenyl ethanol+C2H4O**]**							
				161.05	C6H9O5	**[**Glucose-H2O-H**]**							
				149.05	C5H9O5	**[**Xylose-H**]**							
				131.04	C5H7O4	**[**Xylose-H2O-H**]**							
				119.05	C8H7O1	**[**Phenyl-ethanol-2H-H**]**							
				113.02	C5H5O3								
				101.02	C4H5O3								

7	**furaneol-deoxy-hexosyl-glucose derivative**	13.40	485.1859 ([M-H]-)		C19H33O14								
				436.16	C18H28O12								
				353.12	C17H21O8	[M-H-Pentose]							
				352.12	C17H20O8								
				266.10	C10H18O8								
				205.07	C8H13O6	[Rhamnosyl+C2H4O]							
				163.06	C6H11O5	[Rhamnosyl-H]							
				145.05	C6H9O4	[Rhamnosyl-H-H2O]							
				143.04	C6H7O4	[Rhamnosyl-H-H2O-2H]							
				127.04	C6H7O3	[Furaneol-H]							
				125.02	C6H5O3	[Furaneol-H-2H]							
				103.04	C4H7O3								
8	**trans-linalool-oxide_hexosyl-glucoside**	14.80	539.235 ([M+FA]-)		C23H39O14								
				493.23	C22H37O12	[M-FA-H]							
				331.18	C16H27O7	[M-FA-H-Glucoside]							
				179.06	C6H11O6	[Glucose-H]							
				163.06	C6H11O5	[Glucose-H-O]							
				161.05	C6H9O5	[Glucose-H2O-H]							
				145.05	C6H9O4	[Glucose-H2O-O-H]							
				119.04	C4H7O4								
				103.04	C4H7O3								
				101.02	C4H5O3								
9	**trans-hydroxy-linalool-rhamnosyl-glucoside**	19.90	523.2399 ([M+FA]-)		C23H40O13								
				477.24	C22H38O11	[M-FA-H]							
				331.18	C16H27O7	[M-FA-H-Rhamosyl]							
				247.08	C10H15O7	M-FA-H-trans-linalool-CH2O]							
				205.07	C8H13O6	[Rhamnosyl+C2H4O]							
				163.06	C6H11O5	[Rhamnose-H]							

				161.05	C6H9O5	[Glucose-H-H2O]							
				145.05	C6H9O4	[Rhamnose-H-H2O]							
				143.04	C6H7O4								
				113.02	C5H5O3								
				103.04	C4H7O3								
				101.02	C4H5O3								
10	**cis-hydroxy-linalool-pentosyl-glucoside**	20.11	463.218 ([M-H]-)		C21H35O11		509.225 ([M+FA]-)		C22H37O13				
				331.18	C16H27O7	[M-H-Pentosyl]		463.22	C21H35O11	[M-FA-H]			
				233.07	C9H13O7	[Glucose+Pentosyl fragment]		331.18	C16H27O7	[M-H-FA-Pentosyl]			
				161.05	C6H19O5	[Glucose-H-H2O]		161.05	C6H10O5	[Glucose-H2O-H]			
				149.05	C5H9O5	[Pentose-H]		149.05	C5H9O5				
				143.04	C6H8O4			101.06	C5H9O2				
				131.04	C5H9O4	[Pentose-H2O-H]							
				119.04	C4H8O4								
				113.02	C5H6O3								
				101.02	C4H6O3								
11	**Terpenol_pentosyl-glucoside**	21.00	447.223 ([M-H]-)		C21H36O10		493.22 ([M+FA]-)		C22H38O12		471.22 ([M+Na]+)		C21H36O10Na
				315.18	C16H27O6	[M-H-Pentose]		316.16	C16H28O6			311.10	C11H19O10
				233.07	C9H13O7	[Glucose-Pentose fragment-H]		315.18	C16H27O6	[M-H-FA-Pentose]		309.17	C17H25O5
				191.06	C7H11O6	[Pentose+C2H4O-H]		191.07	C7H11O6	[Pentose+C2H4O-H]		293.17	C17H25O4
				161.05	C6H9O5	[Glucose-H2O-H]		179.07	C6H11O6	[Glucose-H]		229.07	C10H13O6
				159.03	C6H7O5			161.05	C6H9O5	[Glucose-H2O-H]		203.06	C8H11O6
				149.05	C5H9O5	[Pentose-H]		159.04	C6H7O5			201.08	C9H13O5
				143.04	C6H7O4			149.05	C5H9O5	[Pentose-H]		193.16	C13H21O1
				131.04	C5H7O4	[Pentose-H-H2O]		143.04	C6H7O4			163.06	C6H11O5
				119.04	C4H7O4			131.04	C5H7O4	[Pentose-H-H2O]		157.05	C7H9O4
				113.02	C5H5O3			119.04	C4H7O4			155.03	C7H7O4
				101.02	C4H5O3			113.03	C5H5O3				
12	**Terpenol_Rhamnosyl-glucoside**	21.30	461.2388 ([M-H]-)		C22H37O10			101.04	C4H5O3				

				315.18	C16H27O6	[M-H-Rhamnose]							
				205.07	C8H13O6	[Rhamnosyl+C2H4O]							
				163.06	C6H11O5	[Rhamnose-H]							
				161.05	C6H9O5	[Glucose-H2O-H]							
				113.02	C5H5O3								
				103.04	C4H7O3								
				101.02	C4H5O3								
13	**Geranic acid_pentosyl-glucoside**	20.80	507.2224 ([M+FA]-)		C22H35O13								
				461.20	C21H33O11	[M-FA-H]							
				191.06	C7H11O6	[Pentose+C2H4O-H]							
				167.11	C10H15O2	[Geranic acid-H]							
				149.05	C5H9O5								
				131.04	C5H7O4	[Pentose-H-H2O]							
				125.02	C6H5O3								
				113.02	C5H5O3								
				101.02	C4H5O3								
14	**Geranic acid_rhamnosyl-glucoside**	21.05	475.22 ([M-H]-)		C22H35O11		521.2217 ([M+FA]-)		C23H37O13				
				167.12	C10H15O2	[Geranic acid-H]		475.22	C22H35O11				
				103.00	C4H7O3			307.10	C12H19O9				
				101.03	C4H5O3			247.08	C10H15O7	Glucose+Rhamnose fragment]			
								205.07	C8H13O6	[Rhamnosyl+C2H4O]			
								167.11	C10H15O2	[Geranic acid-H]	[Geranic acid-H]		
								163.06	C6H11O5	[Rhamnose-H]			
								145.05	C6H9O4	[Rhamnose-H-H2O]			
								143.04	C6H7O4				
								103.04	C4H7O3				
								101.02	C4H5O3				
15	**vomifolyl-glucoside**	13.90	431.19 ([M+FA]-)		C20H32O10								

				385.19	C19H31O8	[M-H-FA]							
				223.13	C13H19O3	[Vomifoliol-H]							
				205.12	C13H17O2	[Vomifoliol-H-H2O]							
				179.06	C6H11O6	[Glucose-H]							
				161.045	C6H9O5	[Glucose-H-H2O]							
				153.09	C9H13O2	[Vomifoliol-no side chain]							
				152.08	C9H13O2								
				113.02	C5H6O3 -H								
				101.02	C4H5O3								

Table 4: The MS/MS spectra of the 15 peaks described in the section 5.3.5. Some of the precursors were adducts or ions from both positive and negative ionization modes representing the same peak and are listed in the same lines.

5.3.6 Un-correlating putative identifications

Ten MS/MS spectra corresponding to 17 peaks were putatively identified as Terpendiol conjugates. Nevertheless, their peaks area did not show unambiguous hydrolysis and they did not correlate with any of the remaining Terpendiols measured in the GC-MS analysis, so they remained unassigned. One peak was showing a fragmentation pattern corresponding to Vomifoliol-Pnetosyl-glucoside but it did not show any hydrolysis. One further peak was putatively identified as oxo-alpha-ionol_glucoside, but did not correlate with the 2 forms of oxo-alpha-ionol found using GC/MS. Two peaks were putatively identified as Hydroxy-Citronellol_Pentosyl-Glucoside, but they did not show any correlation with Hydroxy-Citronellol quantification. One peak was putatively identified as HomoVanillyl-alcohol-Pentosyl-Glucoside, but no correlations were found with HomoVanillyl-alcohol. One peak was putatively identified as Hotrienol-pentosyl-glucoside, but again no correlation with the GC/MS data of the bound fraction after hydrolysis was found. For all the latter 4 peaks described, other unidentified/unexpected conjugates may contribute to the accumulation of the volatile form after enzymatic hydrolysis.

One peak was identified as putative hydroxy-Geranic acid-Rhamnosyl-Glucoside, on the basis of the presence of a marker peak (183.0999, Image 7), already reported in another work as putative conjugated hydroxy-geranic acid (Yang et al. 2011). Three different peaks were identified as putative Terpentriol-Rhamnosyl-Glucosides, mostly because they showed clear fragmentation spectra of rhamnosyl-glucoside and a peak of 185.1170 (Image 8), corresponding to a Terpentriol-H. Both Terpentriol and hydroxy-geranic acid are expected to have a very high boiling point and were not expected to be seen in GC-MS. The MS/MS spectra of all the compounds described in this section are in the supplementary table 4 attached at the bottom of the thesis.

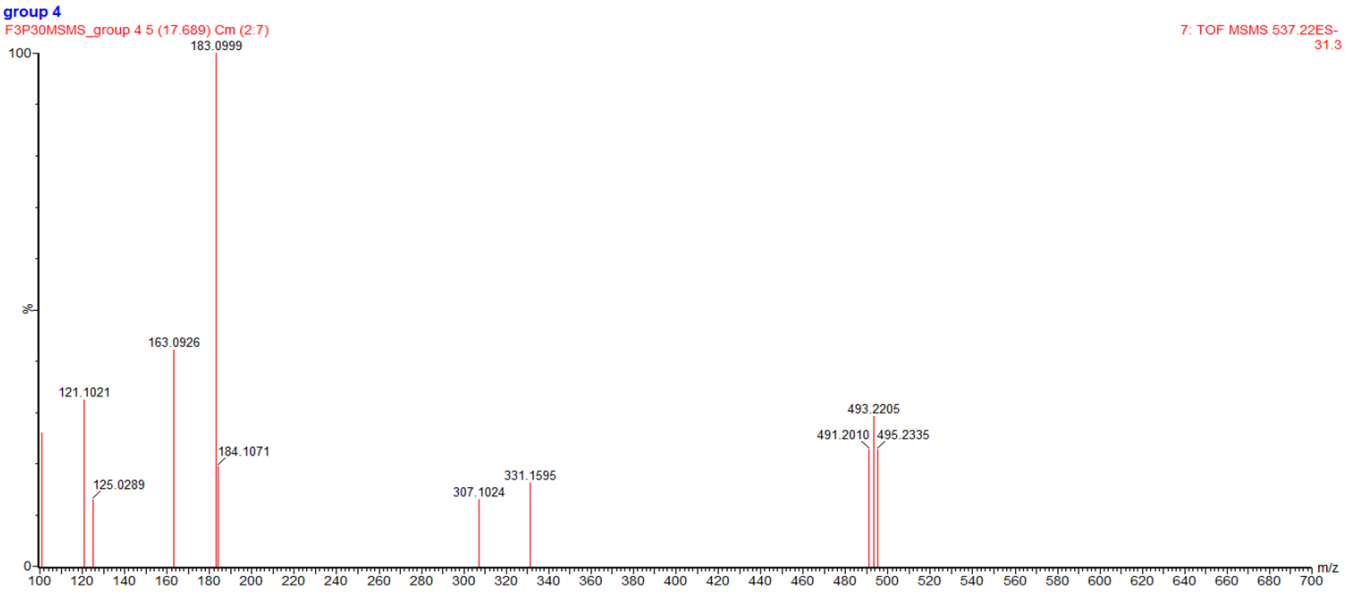


Image 7: MS/MS spectrum of the putative hydroxy-geranic acid_rhamnosyl-glucoside peak (FA adduct). Parent ion 537.2205

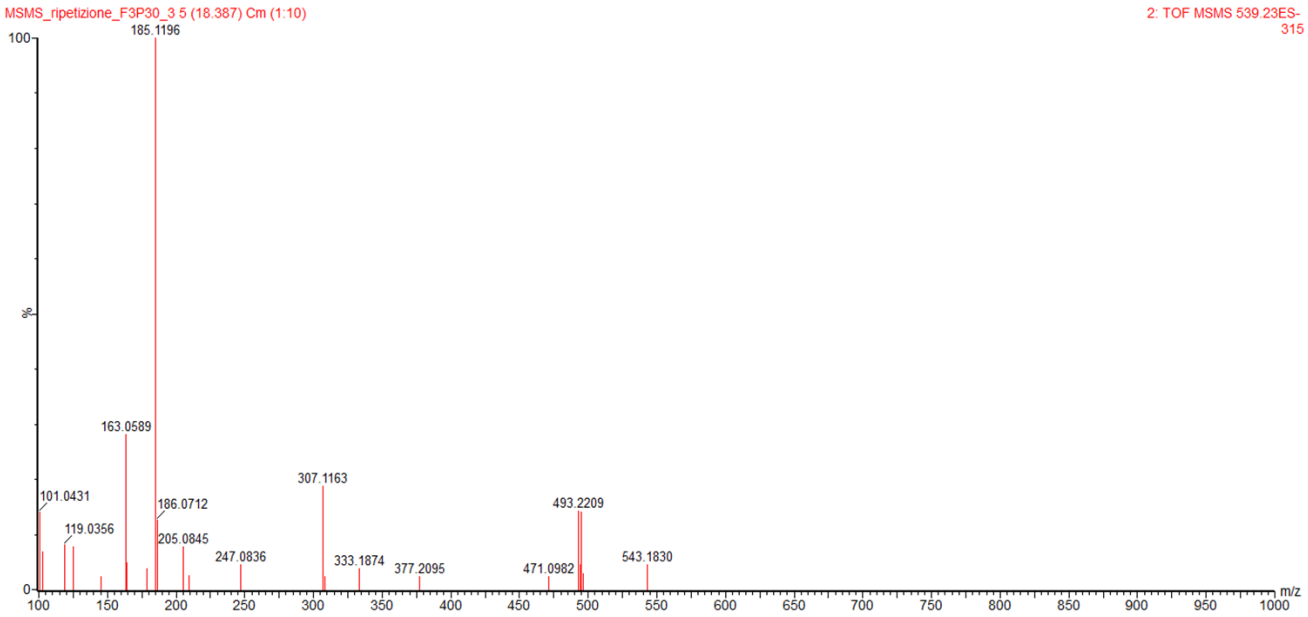


Image 8: MS/MS spectrum of the putative terpentriol_rhamnosyl-glucoside (FA adduct). Parent ion: 539.235

5.3.7 AR2000 enzyme efficiency: post-experimental considerations

The Rapidase AR2000 (Oenobrands) was used for the enzymatic hydrolysis of the glyco-conjugated volatiles. The choice was made based upon several previous studies (Wightman et al. 1997; Baek et al. 1999; Schneider et al. 2004; Vrhovsek et al. 2014). A "golden hydrolysis procedure" has

still to be established, and any choice has some advantages and disadvantages. We ruled out the hydrolysis with strong acids, since it is known to produce several artefacts and because it is not similar to the hydrolysis happening in wine, where the pH is over 3. The enzyme-based strategy with AR2000 was chosen as a milder approach. In a recent study by Flamini et al. (2014) demonstrated the potential of AR2000 for the complete hydrolysis of grape monoterpene glycosides. Nevertheless, we found that AR2000 had a non-specific effect on releasing glycosylated polyphenols (Image 9), as has also been reported in previous papers (Wightman et al., 1997). Furthermore, the data suggests that not all the possible precursors may have been hydrolysed by AR2000, as shown in Image 3; the three peaks before hydrolysis are three different Benzyl-pentosyl-glucosides: after hydrolysis the ion currents of two of them disappear, while one decreases but does not disappear. This indicates that AR2000 has a specific effect on the hydrolysis of some glycosylated precursors and could have a minor effect (or no effect) on other glycosides. The data suggests that the potential bound aroma released, measured using GC/MS after enzymatic hydrolysis, is only part of the overall potential bound aroma of our cultivars. To our knowledge, this is the first time that the efficiency of hydrolysis on several different classes of conjugates has been thoroughly tested against their quantitative release measured by GC-MS.

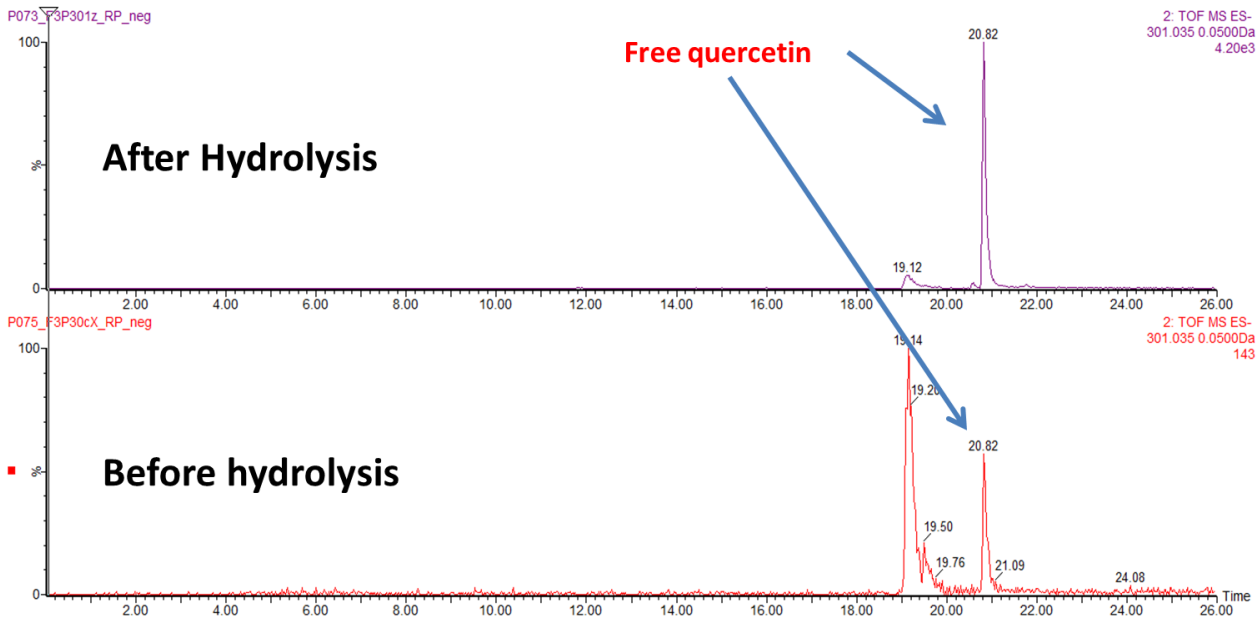


Image 9: Formation of free Quercetin after enzymatic hydrolysis.

5.4 Conclusion

The aim of the work described in chapter 4 was to putatively identify the Glyco-conjugated volatiles in the intact form using LC-MS analysis. Through features filtering (section 5.3.1), MS/MS analysis (section 5.3.3) and correlation of the LC-MS data with GC-MS data (section 5.3.5), I was able to identify 15 previously unreported glyco-conjugated volatiles. Moreover another 17 different glyco-conjugated volatiles have been putatively identified.

Even if some of the compounds showed to be unique to some grape varieties, the differences between the samples have not been studied in this work, because a better planned experiment to explore varietal differences has been performed and it is described in chapter 7. This work must be considered as a step forward in the exploration of the grape metabolome and as a basic work to the comparative analysis of grape species described in chapter 7.

References Chapter 5

1. *Alfassi, Z. B. (2004). On the normalization of a mass spectrum for comparison of two spectra. Journal of the American Society for Mass Spectrometry, 15(3), 385–7. doi:10.1016/j.jasms.2003.11.008*
2. *Arapitsas, P., Speri, G., Angeli, A., Perenzoni, D., & Mattivi, F. (2014). The influence of storage on the "chemical age" of red wines. Metabolomics, 10(5), 816–832. doi:10.1007/s11306-014-0638-x*
3. *Baek, H. H., & Cadwallader, K. R. (1999). Contribution of Free and Glycosidically Bound Volatile Compounds to the Aroma of Muscadine Grape Juice. Journal of Food Science, 64(3), 441–444. doi:10.1111/j.1365-2621.1999.tb15059.x*
4. *Baldwin, I. T., Halitschke, R., Paschold, A., Dahl, C. C. Von, & Preston, C. A. (2006). Volatile Signaling in Plant-Plant in the Genomics Era, (February), 812–816.*
5. *Boido, E. D., LLoret, A. D., Medina, K. A. M., & Farin, L. A. (2003). Aroma Composition of Vitis vinifera Cv . Tannat : the Typical Red Wine from Uruguay, 5408–5413.*
6. *Ciani, M., Comitini, F., Mannazzu, I., & Domizio, P. (2010). Controlled mixed culture fermentation: a new perspective on the use of non-Saccharomyces yeasts in winemaking. FEMS Yeast Research, 10(2), 123–33. doi:10.1111/j.1567-1364.2009.00579.x*
7. *Esti, M., & Tamborra, P. (2006). Influence of winemaking techniques on aroma precursors. Analytica Chimica Acta, 563(1-2), 173–179. doi:10.1016/j.aca.2005.12.025*
8. *Fernández-González, M., & Di Stefano, R. (2004). Fractionation of glycoside aroma precursors in neutral grapes. Hydrolysis and conversion by Saccharomyces cerevisiae. LWT - Food Science and Technology, 37(4), 467–473. doi:10.1016/j.lwt.2003.11.003*
9. *Flamini, R., De Rosso, M., Panighel, A., Dalla Vedova, A., De Marchi, F., & Bavaresco, L. (2014). Profiling of grape monoterpene glycosides (aroma precursors) by ultra-high performance-liquid chromatography-high resolution mass spectrometry (UHPLC/QTOF). Journal of Mass Spectrometry : JMS, 49(12), 1214–22. doi:10.1002/jms.3441*
10. *Franceschi, P., Mylonas, R., Shahaf, N., Scholz, M., Arapitsas, P., Masuero, D., … Wehrens, R. (2014). MetaDB a Data Processing Workflow in Untargeted MS-Based Metabolomics Experiments. Frontiers in Bioengineering and Biotechnology, 2(December), 72. doi:10.3389/fbioe.2014.00072*

11. *Ghaste M.; Narduzzi L.; Carlin S.; Vrhovsek U.; Shulaev V.; Mattivi F, (2015) Chemical Composition of Volatile Aroma Metabolites and their Glycosylated Precursors Uniquely Differentiates Individual Grape Cultivars, DOI: 10.1016/j.foodchem.2015.04.056*
12. *Kuhl, C., Tautenhahn, R., Bo, C., Larson, T. R., & Neumann, S. (2012). CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets. Analytical Chemistry.*
13. *Little, J. L. (1999). Artifacts in trimethylsilyl derivatization reactions and ways to avoid them. Journal of Chromatography A, 844(1-2), 1–22. doi:10.1016/S0021-9673(99)00267-8*
14. *Lund, S. T., & Bohlmann, J. (2006). The Molecular Basis for Wine Grape, 311(February), 804–806.*
15. *Maicas, S., & Mateo, J. J. (2005). Hydrolysis of terpenyl glycosides in grape juice and other fruit juices: a review. Applied Microbiology and Biotechnology, 67(3), 322–35. doi:10.1007/s00253-004-1806-0*
16. *Robinson, a. L., Boss, P. K., Solomon, P. S., Trengove, R. D., Heymann, H., & Ebeler, S. E. (2013). Origins of Grape and Wine Aroma. Part 1. Chemical Components and Viticultural Impacts. American Journal of Enology and Viticulture, 65(1), 1–24. doi:10.5344/ajev.2013.12070*
17. *Sarry, J., & Gunata, Z. (2004). Plant and microbial glycoside hydrolases: Volatile release from glycosidic aroma precursors. Food Chemistry, 87(4), 509–521. doi:10.1016/j.foodchem.2004.01.003*
18. *Schievano, E., D'Ambrosio, M., Mazzaretto, I., Ferrarini, R., Magno, F., Mammi, S., & Favaro, G. (2013). Identification of wine aroma precursors in Moscato Giallo grape juice: a nuclear magnetic resonance and liquid chromatography-mass spectrometry tandem study. Talanta, 116, 841–51. doi:10.1016/j.talanta.2013.07.049*
19. *Schneider, R., Razungles, a, Augier, C., & Baumes, R. (2001). Monoterpenic and norisoprenoidic glycoconjugates of Vitis vinifera L. cv. Melon B. as precursors of odorants in Muscadet wines. Journal of Chromatography A, 936(1-2), 145–157. doi:10.1016/S0021-9673(01)01150-5*
20. *Sleno, L. (2012). The use of mass defect in modern mass spectrometry. Journal of Mass Spectrometry, 47(2), 226–236. doi:10.1002/jms.2953*

21. *Smith, C. a, Want, E. J., O'Maille, G., Abagyan, R., & Siuzdak, G. (2006). XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Analytical Chemistry, 78(3), 779–87. doi:10.1021/ac051437y*
22. *Tikunov, Y. M., de Vos, R. C. H., González Paramás, A. M. X., Hall, R. D., & Bovy, A. G. (2010). A role for differential glycoconjugation in the emission of phenylpropanoid volatiles from tomato fruit discovered using a metabolic data fusion approach. Plant Physiology, 152(1), 55–70. doi:10.1104/pp.109.146670*
23. *Vrhovsek, U., Lotti, C., Masuero, D., Carlin, S., Weingart, G., & Mattivi, F. (2014). Quantitative metabolic profiling of grape, apple and raspberry volatile compounds (VOCs) using a GC/MS/MS method. Journal of Chromatography. B, Analytical Technologies in the Biomedical and Life Sciences, 966, 132–9. doi:10.1016/j.jchromb.2014.01.009*
24. *Wang, J., & Hou, B. (2008). Glycosyltransferases: key players involved in the modification of plant secondary metabolites. Frontiers of Biology in China, 4(1), 39–46. doi:10.1007/s11515-008-0111-1*
25. *Wightman, J. D., & Wrolstad, R. E. (1997). β-glucosidase Activity in Juice-Processing Enzymes Based on Anthocyanin Analysis, 61(3), 544–548.*
26. *Wolf, S., Schmidt, S., Müller-Hannemann, M., & Neumann, S. (2010). In silico fragmentation for computer assisted identification of metabolite mass spectra. BMC Bioinformatics, 11, 148.*
27. *Yang, T., Stoopen, G., Yalpani, N., Vervoort, J., de Vos, R., Voster, A., … Jongsma, M. a. (2011). Metabolic engineering of geranic acid in maize to achieve fungal resistance is compromised by novel glycosylation patterns. Metabolic Engineering, 13(4), 414–25. doi:10.1016/j.ymben.2011.01.011*

6. The compound characteristics comparison method (CCC): current identification strategies, method development and its integration with state-of-the-art methodologies.

The main bottleneck of the untargeted metabolomics analysis is the identification of the m/z spectra obtained from the LC-MS analysis. The putative number of metabolites creating MS spectra shown in the chromatogram is over one thousand, and their identification is very difficult. There are several problems in the identification of the compounds, which have been described in chapter 4. In this "proof of principles" study, I tried to overcome the limits of the analysis, developing a new method to establish a regression model between different metabolites, trying to demonstrate that metabolites showing similar features have also a similar structure, and a comparable chemical class. The coupling of this information with the MS/MS spectra helps to individuate the correct chemical formula and the most putative structure.

This project has been carried out by me, under the supervision of Dr. Fulvio Mattivi and Dr. Pietro Franceschi. In this project, isotopic intensities extraction has been obtained using the script developed by Dr. Jan Stanstrup, to which I give credits in the description. R scripts development has been also supervised by Dr. Jan Stanstrup.

6.1 Introduction

In current untargeted metabolomics experiments, the identification of the metabolites is a key step to give a biological meaning to the experiment. Indeed, the identification of all the metabolites that show a different pattern between the control group (group A) and the treatment group (group B) would be the perfect condition in a metabolomics experiment. Unfortunately, this is not the case: identification of all the metabolites is far to be achieved, and if automating spectral matching between samples and chemical standards is slowly becoming a reality (Wehrens et al. 2014), the identification of the unknown metabolites is still a time-consuming and error-prone process.

According to the metabolomics society, there are four accepted levels of identification, reported by Sumner et al. (2007). MSI level one corresponds to complete identification through comparison of two or more orthogonal characteristics (e.g. RT and MS spectra). MSI level 2 and 3 are putative identification of the compound name or the compound class respectively, while MSI level four is the raw RT and m/z signature. If the MSI level four sounds unpleasant to be presented in a scientific paper, achieving the MSI level 3 based only on the molecular ion (without having any further information) is already hard. There are four steps that allow achieving, or at least improving, the identification of a

compound: Chemical Formula calculation, MS/MS spectra interpretation, RT prediction and biosynthetic pathway comparison.

The easiest and quickest way to achieve the MSI level 3 is to obtain the Chemical Formula. In the 1950, a researcher named Van Krevelen, demonstrated that compounds with a similar H/C and O/C ratios are more likely part of the same chemical class (Van Krevelen, 1950). It displayed the results in a bidimensional plot, and dividing the plot in different parts, was possible to separate the dots in different chemical classes. Furthermore, the plot allows describing the reactions undergoing in the samples (Werner et al. 2008). Three-dimensional and multidimensional Van Krevelen diagrams exist (Wu et al. 2004), adding as further dimensions the ion intensity, the N/C and S/C ratios, etcetera. This diagram is valid for every matrix, and found large application in the oil and organic matter research. The main limit of the diagram is that it needs as input all the chemical formulas of all the signals obtained in the mass spec, which is achievable only with ultra-high resolution mass spectrometers (like FT-ICR-MS), and only up to 7-800 Daltons, with a limited number of elements. Formula calculation is a factor dependent from the accuracy in the determination of the m/z and isotopic pattern of the metabolites. In the instrument used in my experiment ("Synapt") mass error can reach 30 ppm (Shahaf et al. 2013), and is in average around five ppm. The error in the detection of the isotopic clusters is usually low in the TOF instruments, but can be very high, when disturbing ions are detected at the same RT of the analytes of interest (Thurman & Ferrer, 2010). Therefore, it is not an easy task to calculate.

MS/MS spectra acquisition and interpretation is probably the older and more reliable method to improve the identification of a compound. In the MS/MS spectrum, every ion signal in the spectrum represents a fragment of the selected parent ion. Because fragmentation is not a random process but follows common rules (at least in the same instruments), the MS/MS spectra can be compared with experimental spectra from databases (Massbank, Metlin)and assigned to putative structures. It also can be queried against in silico-fragmentation simulators like MetFrag, Metfusion and Sirius (Wolf et al. 2010, Gerlich et al. 2013, and Rasche et al. 2010).

On the other hand, MS/MS spectra acquisition is not a straightforward process: in Q-TOF, instruments there are mostly two ways to perform MS/MS spectral acquisition: 1) selecting the precursor ions (so-called MS/MS analysis) and 2) fragmenting all the precursors together, (MS^e) using a ramp of collision energy and rebuilding the ion relationships with dedicated software. The first method is the one used in this thesis, and its limit (the possibility of selecting multiple distinct parents) has been already discussed in chapter 2. The second method works very well with very intense ions, but with medium to low intensity, ions the spectral reconstruction might become confuse.

The RT time prediction is based on the construction of regression model between the calculated physico-chemical properties of a set of compounds and their retention time (Boswell et al. 2011, Creek et al. 2011). If the model is stable, the retention time detected for an unknown metabolite can be used to filter out all the putative structures unlike to have such RT. A second approach would be the injection of the same standard set in different LC-MS method, and a successive alignment of the obtained RT to allow to predict the RT of the compounds inject in one of the method, with an acceptable confidence interval (Boswell et al. 2011, Stanstrup et al. 2013).

The previous three steps allow the analyst to have one (or few) putative structures for an unknown biomarker. The predicted structures are further integrated in the metabolic pathways of the organism under analysis, to confirm their structure and explain their biological meaning. Probabilistic methods to assign the correct formula has been developed by Rogers et al. (2009), while Bayesian methods to assign structures based on spectrometric data coupled with pathway analysis has been established by Silva et al. (2014).

The identification of the compounds is based on these four steps that are generally considered as independent filters, used one after another. Being used as independent "filters", the measurement errors are affecting very much their filtering properties; big measurement error means broad filter or no-filtering at all. On the other hand, these "filters" are describing a unique entity, the chemical structure.

In literature, there are not methods that use the four "filters" as a whole. The interpolation of the information of the four "filters" together may overcome the effects that the error has on each of the "filters", improving the confidence in the identification. However, how is that possible to combine four independent or semi-dependent characteristics? To what we compare eventually, the results obtained from a combined analysis of their characteristics?

The answers to these two questions are the aim of this project and will be described and discussed in this "proof of principles" study.

6.2 Basic concepts

6.2.1 Multivariate statistics to predict the model performance

In the introduction, we underlined the necessity of the treatment of the numerous characteristics of a compound as a whole. Many physico-chemical characteristics of the compounds are inter-correlated and in some cases might have a nonlinear regression with the characteristics that we are trying to predict (the chemical structure and the chemical formula). This means that we need to apply a

regression model that is able to manage correlated predictors and eventually transform the values to have a linear model. To understand if this step was feasible, I subdivided the standards of our database according to their chemical classes and I used it as training set for my model (the list of the standard used is the same of Shahaf et al. 2013). The compounds have been subdivided in the following classes:

1) Organic acids,
2) Amines,
3) Amino acids,
4) Phenols (compounds containing a phenolic group),
5) Polyphenols (compounds containing more than a phenolic group),
6) Aromatic amines (compounds containing phenolic groups and amino groups)
7) Polar lipids (lipids having a hydrophilic substituent)
8) Apolar lipids
9) Sugars

To predict the chemical classes of the compounds a very useful method might be the Partial Least Squares discriminant analysis (PLS-da, Stahle & Wold, 1987). In this classification method, two inputs of data must be given to build the model: the X predictor's matrix (independent variables) and the Y classification matrix (dependent variables). In the next section, which prediction variables and why these variables have been chosen will be explained.

6.2.1.1 The predictors matrix (X matrix)

For predictors, I intend the X independent variables that are measurable in a common high-resolution LC-MS instrument. The most common predictors of a chemical structure are the 1) Retention Time (**RT**) and the 2) Monoisotopic Mass (**MM**). In reverse phase, the **RT** has a direct correlation with the hydrophobicity of the compound, which intrinsically depends on the chemical structure of the compound. Aromatic groups, aliphatic chains, and methilic substituents tend to increase hydrophobicity of a compound, while amino groups, acidic and alcoholic substituents increase the hydrophilicity. Being the RT a separation parameter, is obvious that different classes have different RT ranges as shown in the box plot of the image 1.

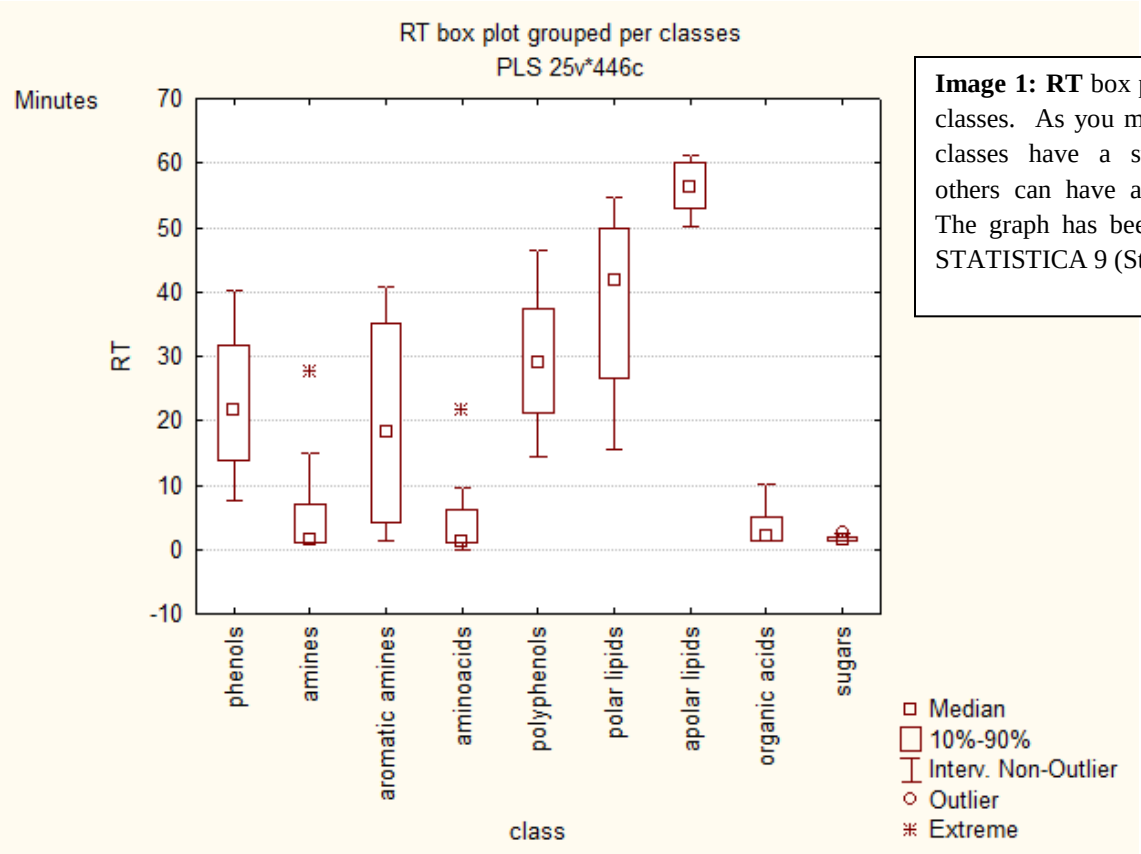


Image 1: RT box plot grouped per classes. As you may notice, some classes have a strict RT while others can have a broader value. The graph has been realized with STATISTICA 9 (StatSoft).

The 2) Monoisotopic Mass (**MM**) is calculated as the exact molecular weight of the most common isotopes of the ions present in the structure. In my model, it can be considered as a predictor of the complexity of the structure: higher the m/z, higher is the number of atoms and the number of subunits composing the chemical structure. In the Image 2 is shown the graph for this parameter of our in-house standard database; only few chemical classes have compounds with MM above 450 Dalton. The third parameter used is the 3) **odd mass**. This choice is based on the Nitrogen rule: in organic chemistry, at the basal level of valence of the atoms, the only atom having an even mass and odd valence is the Nitrogen. This means that all the structures having an even monoisotopic mass number will have an even number of Nitrogens (including 0), while the ones having an odd monoisotopic mass will have an odd number of Nitrogens in their formulas. So, whenever an odd monoisotopic mass is detected, an odd number of Nitrogens must be expected (one, 3, 5 and so on); obviously, an odd monoisotopic mass produces ions with even m/z. Some exceptions exist to this rule, but they regard only statistically 5% of the cases and few chemical classes (Vessecchi et al. 2007). This parameter helps in understanding the presence/absence of Nitrogen in the studied structures.

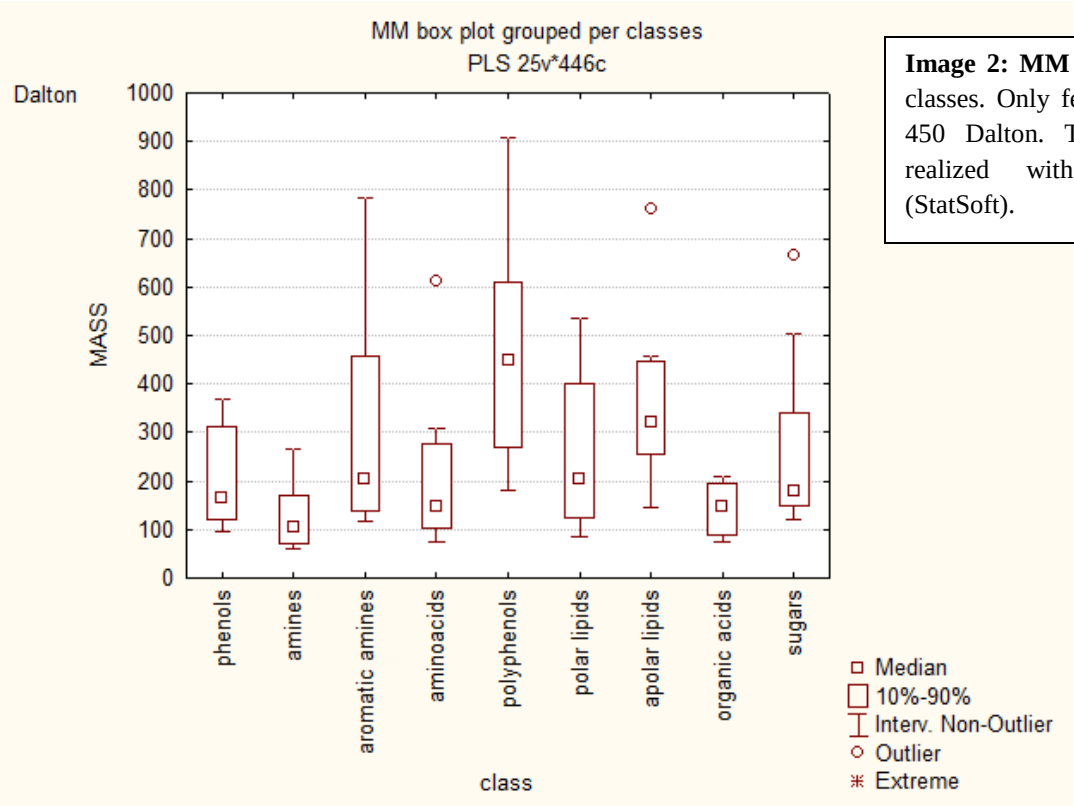


Image 2: MM box plot grouped per classes. Only few classes go beyond 450 Dalton. The graph has been realized with STATISTICA 9 (StatSoft).

Another very important predictor is the 4) mass defect (**md**); for mass defect, is intended the decimal after the **MM** comas different from the zero value. It can be calculated by the formula

$$\boldsymbol{MM - nm = md}$$

With **nm** being the Integer mass (nominal mass chemically speaking). Indeed this value is independent of the number of Carbons in the Formula (because Carbon has 0 mass defect by IUPAC definition), but is a good indicator of the number of ions with positive mass defect (H and N) and negative mass defect (O, P, S). A filter build on the nominal mass and the mass defect has been already proposed by Zhang et al. (2003), to screen the presence of drug in human liquids (urine, blood, serum). In Image 2 is displayed the box plot of our home-build standards database. The **md** range is similar in many classes, except for polar and apolar lipids and for organic acids.

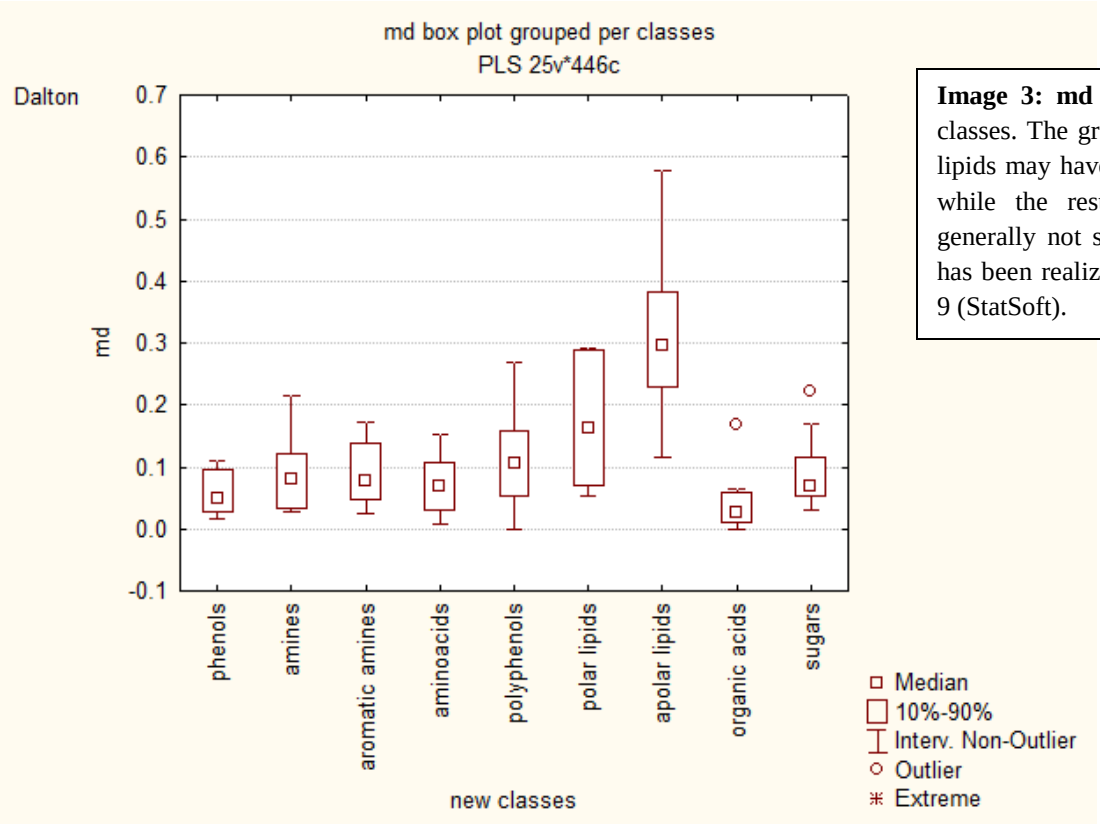


Image 3: md box plot grouped per classes. The graph clearly shows that lipids may have a wide range of **md**, while the rest of the classes are generally not so variable. The graph has been realized with STATISTICA 9 (StatSoft).

Next important predictor is the 5) Relative Mass Defect (**RMD**): it is calculated as the ratio between the mass defect divided by the nominal mass (nm) multiplied per 1000000 (expressed in ppm).

$$\boldsymbol{md/nm * 1000000 = RMD}$$

RMD has been demonstrated to be representative of the oxidation status of soils (Kramer et al. 2001), and it is, in general, a scale to divide different metabolites in different chemical classes (Sleno, 2012). In facts, RMD can be considered as the ratio of (H+N)/(C+O+P+S). According to Sleno (2012), "alkanes have RMD >1000 ppm, membrane lipids and steroids fall within 600 and 1000 ppm, sugars between 300 and 400 ppm, and organic acids with less than 300 ppm". In the image 4, RMD of the different chemical classes of our internal database has been compared. The values reported from Sleno are respected from my classification; furthermore, it is very interesting to observe that some classes have peculiar ranges, like polyphenols (that in our dataset were the most represented class, with 125 distinct compounds). On the other hand, this parameter is useless with amines that have a very broad range.

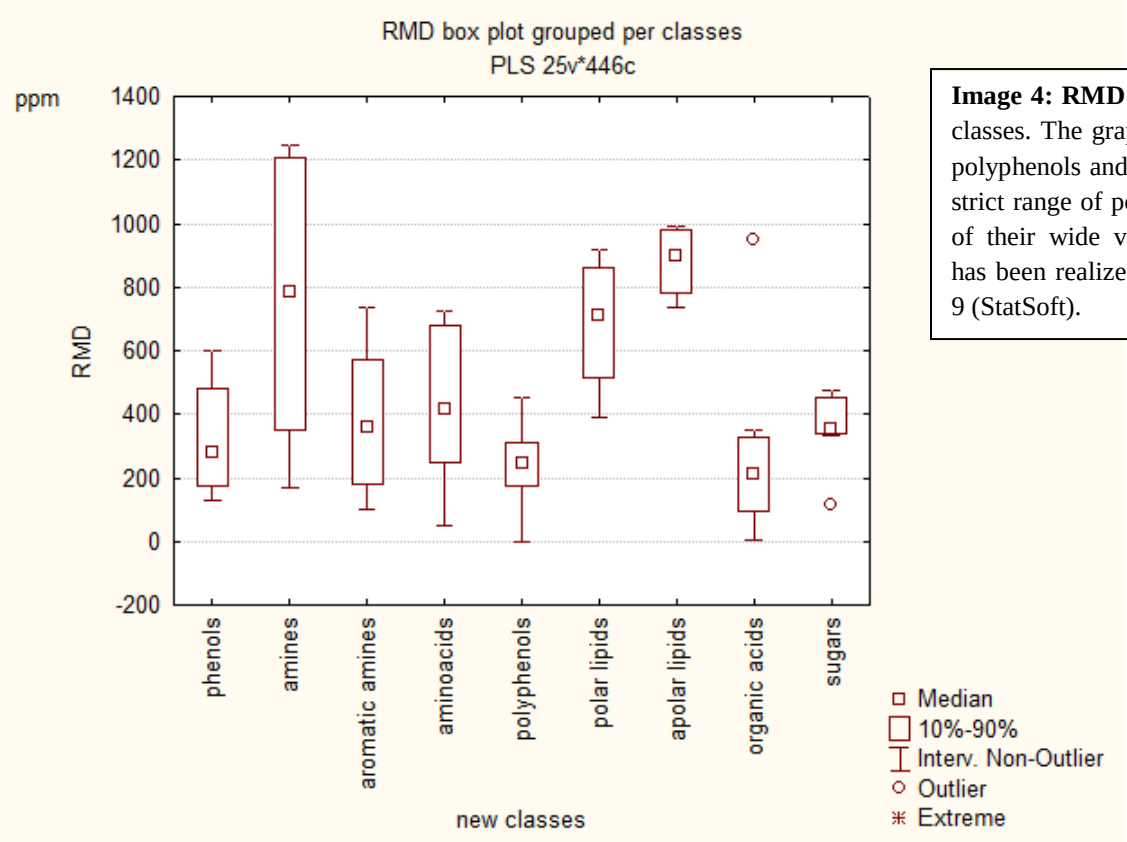


Image 4: RMD box plot grouped per classes. The graph clearly shows that polyphenols and organic acids have a strict range of possible RMD, despite of their wide variability. The graph has been realized with STATISTICA 9 (StatSoft).

The 6) carbon content or number of Carbons (**nofC**) is an extremely important parameter. The amount of Carbons in facts is the main parameter that determines the number of bonds in an organic molecule structure and its inner complexity. In the Image 5, the box plot with the nofC of our internal standard database is displayed; generally, organic acids, amines, and phenols have a limited amount of Carbon in their formula. To calculate the amount of Carbons present in a structure, we need to take in consideration that the isotopic distribution of the Carbon is 98.9% for C12 and 1.1% for C13. Therefore, the ratio between the second isotope and the first is directly correlated with the number of Carbons. The number of Carbons can be calculated building a linear regression curve between the theoretical isotopic ratio between the first isotope and the second isotope and applying this curve to the data. The experimental error in Q-TOF instruments is generally very low with an error in the range of the ± √(correct value). Low molecular weight compounds have low A+1 intensity, but the elemental composition calculations is considered problematic only for masses above 300 Da (Kind & Fiehn, 2006; Knolhoff et al. 2014). The number of Carbons is a direct indicator of the complexity of the

structure and the number of bonds that a structure may have. The construction of a linear regression curve for **nofC** calculation is discussed below.

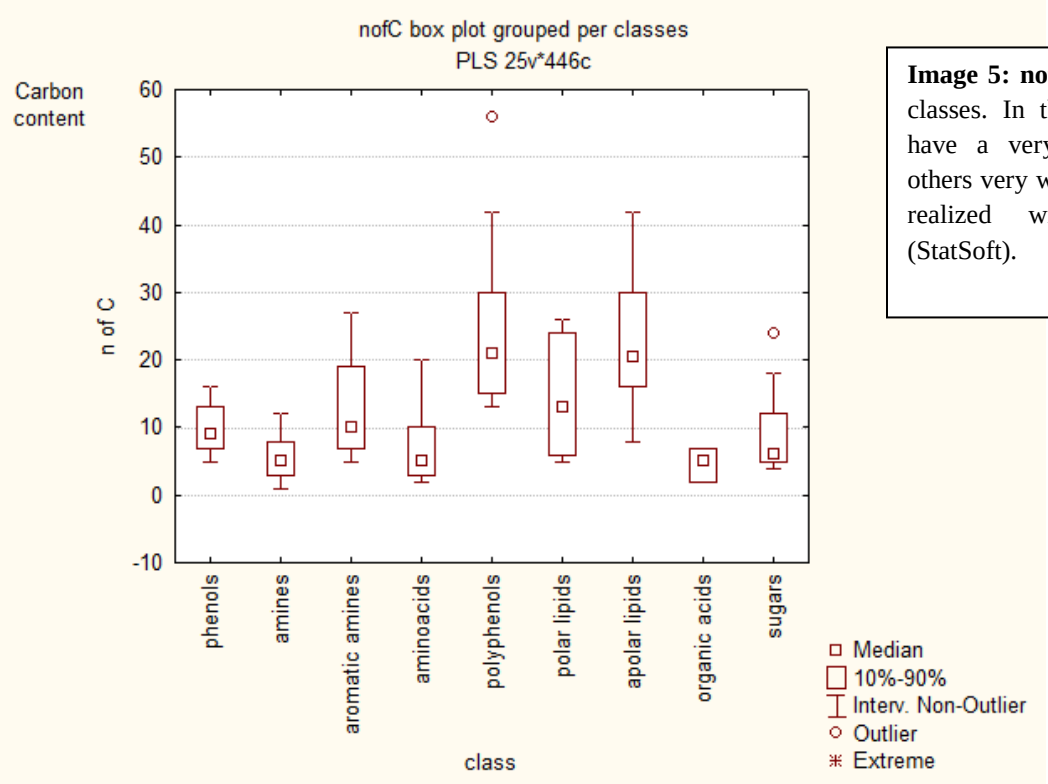


Image 5: nofC box plot grouped per classes. In the graph, some classes have a very narrow range, while others very wide. The graph has been realized with STATISTICA 9 (StatSoft).

The 7) percentage of Carbon mass (**pC**) is simply the fraction of the monoisotopic mass represented from the amount of Carbon content. This value is obtained by the formula

$$(\boldsymbol{nofC} * \mathbf{12})/\boldsymbol{MM} = \boldsymbol{pC}$$

This parameter is indicating somehow the ratio between all the other elements and Carbon, and it is a good indicator of the chemical class of the compounds. For example a compound containing only Carbon and Hydrogen will have a very high **pC**, because the ratio between Hmass/Cmass is 1/12. On the other hand, compounds rich in other elements like Oxygen, Nitrogen, Sulfur and Phosphorous have a very low **pC** (for example acids or sugars). In the image 6 the **pC** of our home-build database is displayed. As shown in the graph, sugars, organic acids and apolar lipids have very narrow ranges, while all the other classes have a wider range. Interestingly, amino acids show a wide range, shifted in comparison to almost all the other classes.

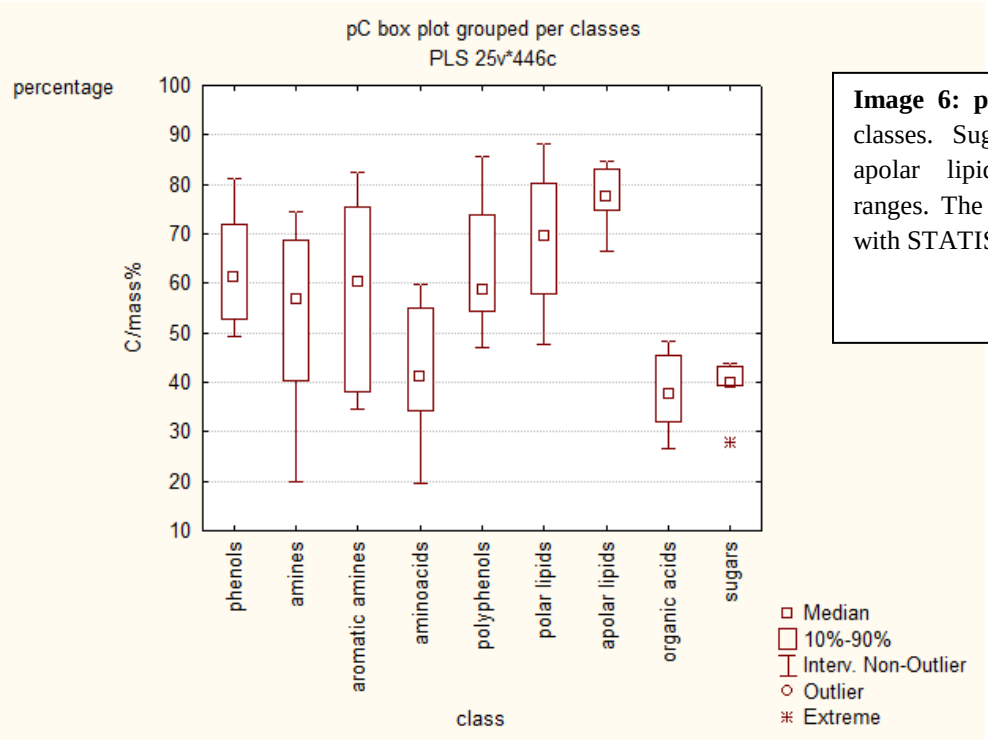


Image 6: pC box plot grouped per classes. Sugars, organic acids and apolar lipids have very narrow ranges. The graph has been realized with STATISTICA 9 (StatSoft).

From the rmass is possible to calculate the 8) residual Relative Mass Defect (**rRMD**), that is calculated as the **RMD** of the residual mass (rmass) after the subtraction of the Carbon mass. This value is calculated like

$$\boldsymbol{MM - (nofC * 12) = rmass}$$

$$\boldsymbol{md(rmass)/nm(rmass) * 1000000 = rRMD}$$

This value is intended to separate classes that have an inconstant elemental ratio the classes with constant elemental ratio. A clear example can be found in image 7. The box plot shows that polyphenols, despite of the class size (125 different metabolites), show a very narrow range, while polar lipids, show a very wide range, but it is almost not overlapping with the polyphenolic one.

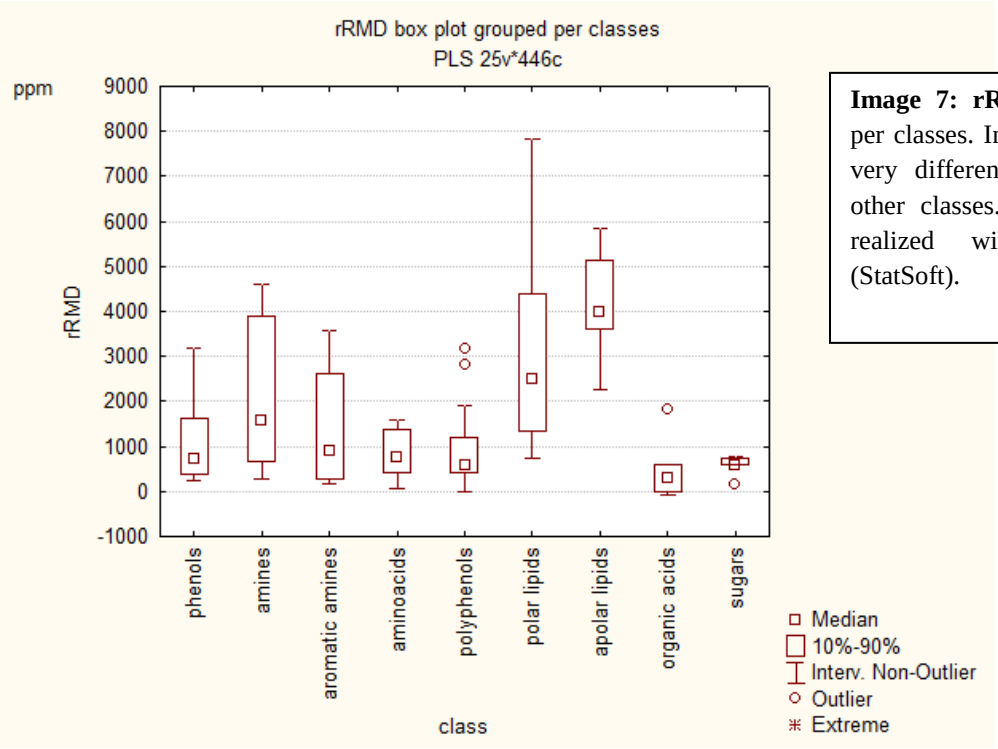
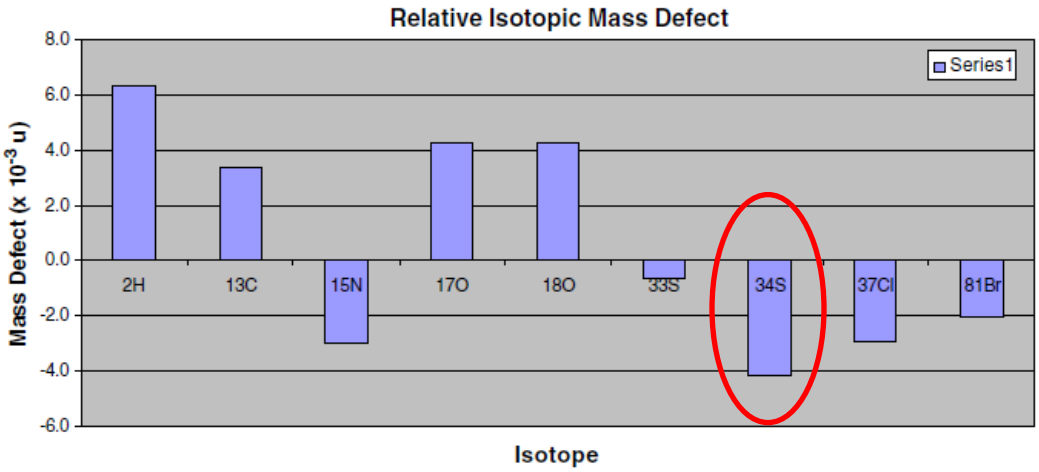


Image 7: rRMD box plot grouped per classes. In the graph, lipids show very different ranges from all the other classes.. The graph has been realized with STATISTICA 9 (StatSoft).

The last parameter used is the 9) third Isotope Mass Defect and Pattern (**tIMDP**), that basically takes in consideration the influence of the third isotope in formula calculations. It is well know that many elements have peculiar isotopic patterns (i.e. Sulfur, Potassium, and Chlorine). The peculiar pattern is helpful to understand when these kinds of compounds are present in the formula or not. If for Potassium, Chlorine, Bromine and others, the intensity shift of the third isotope is quite evident and can be used directly as marker of their presence, this is not the case for Sulfur. Sulfur has a unique isotopic distribution, with 95% of the first isotope, 0.7% of the second isotope and 4.2% of the third isotope. The 4.2% at the third isotope can be confused with high amount of Oxygen in the formula, especially because usually intensity errors are frequent in low intense ions. This often confuses the formula calculations. The **tIMDP** combines the isotopic pattern with the mass defect of the isotopes: the third isotope of a molecule containing Sulfur has a mass defect inferior to the mass defect of the first isotope (Thurman & Ferrer, 2010) . This means that when in the formula there is Sulfur, the third isotope shows a higher intensity and a decreased mass defect. Combining these two informations is possible to assess the presence/absence of Sulfur in the formula.



Sulfur mass	**S(32)** 31.972072	**S(33)** 32.971459	**S(34)** 33.967868	**S(36)** 35.967079
Isotopic distribution	*95.02*	*0.75*	*4.21*	*0.020*

Image 8: In the picture, the isotopic mass defect of the Sulfur34 is underlined. IMD of Sulfur34 is lower than the one from Chlorine and Bromine. The picture is from Thurman & Ferrer, 2010. In the table, Isotopic distribution and masses are displayed. The data displayed is from *http://www.sisweb.com/referenc/source/exactmas.htm*

6.2.2 The responses matrix (Y matrix)

6.2.2.1 Classification approach

The next step of a model building is to choose the responses that we want from the model (the questions we are asking to the model). As shown in the previous section, each of the X independent variables can work as a singular classifier, giving a more or less stringent classification of the different chemical classes. The first approach tested in this work, was to use the potential of the Soft Independent Modeling of Class Analogy approach (SIMCA), coupled with PLS-da analysis to take advantage of the classificative properties of each of the X predictors (Wold, 1976, Wold et al. 1989), to predict the chemical classes of the different metabolites. In the SIMCA approach, the classification of the observations is achieved determining the principal components of the X matrix. The observations are geometrically mapped in a components-driven hyper plane and the class of the observations is assigned according to their Euclidean distance with a confidence interval of 95%, determined through cross-validation (CV) with PLS-da algorithm. This classification method has been demonstrated to be very valuable and versatile, being adaptable to different datasets and different principal component analysis (Branden & Hubert 2005, Bylesio et al. 2007).

The subdivision done by the SIMCA approach is not sharp; if an observations falls in an area between two classes, it can be assigned to both, according to the confidence intervals (previously calculated with CV). This classification method was thought to be used to assign a chemical class to every metabolite according to the X matrix and their respective classes. I developed this part using SIMCA-P+12.0.0 software (Umetrics). First a PCA analysis is performed to see if the components can describe the variability hold in the X matrix. The PCA plot is shown in image 9. The software returned two components, accounting for 68% of the total variability. The Different classes have been colored, to observe their plotting. It is obvious that components can describe the variability present in the X matrix.

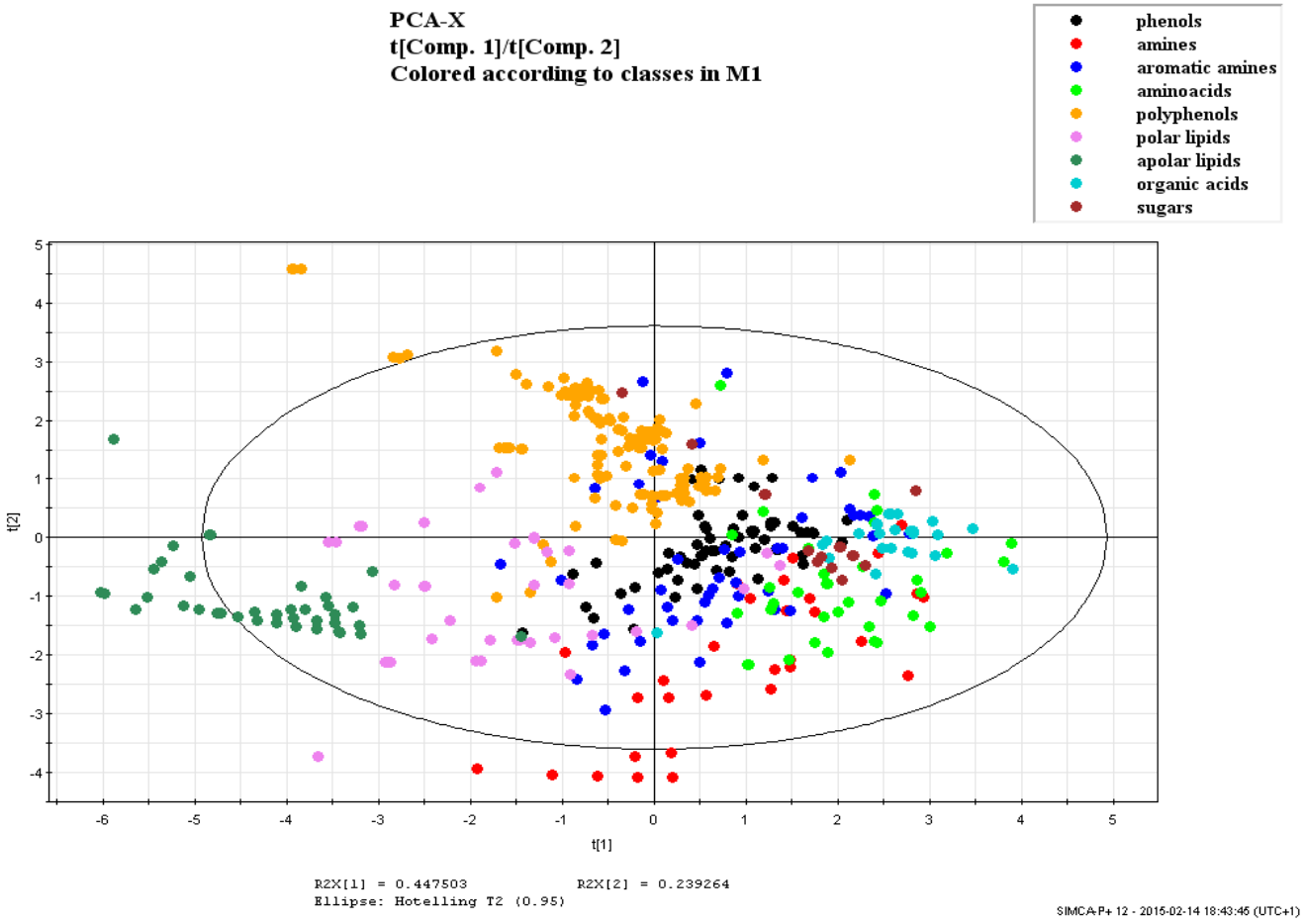


Image 9: A PCA plot of the X matrix of all the 446 compounds from the internal standards database. The first two components account for about 68% of the variability. Some class separated clearly after two components (amines, apolar lipids and polyphenols), while others were overlapping.

The next step was to fit a PLS-da model, to see if the partial least square analysis could predict the classes. Unfortunately this was not the case. In the image 10, the histograms indicate the total amount of variability explained by the model after fitting the next component. The model had eight components, and in total they could account for the 36% of the total variability of classes as shown in the image 10.

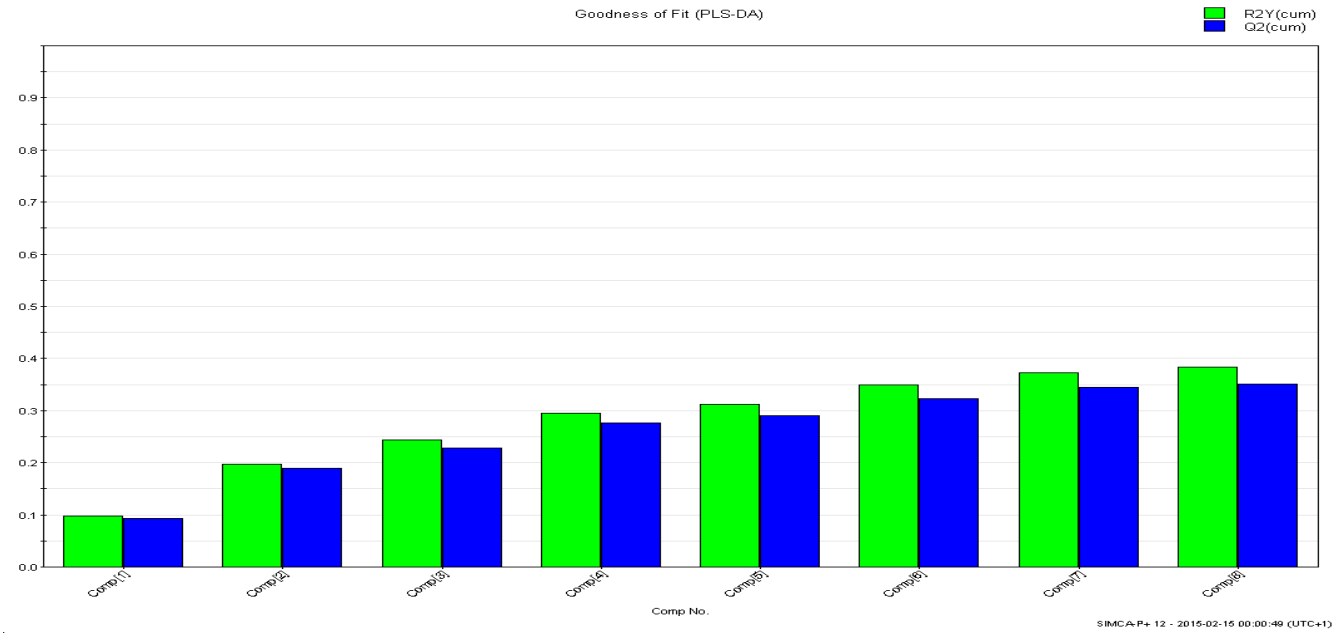


Image 10: The cumulative amount of Y variance explained (classes) from the X matrix after eight components. A good prediction is usually above the 0.5 value.

To better evaluate the classificative properties of the model and try to understand why the fit was not so promising I performed the misclassification analysis. In the misclassification analysis, using CV, all the compounds present in the model are assigned to one class, according to the highest classification value obtained by the model. The results are displayed in table 1.

	1	2	3	4	5	6	7	8	9	10	11	12	13
1		Members	Correct	phenols	amines	aromatic amines	aminoacids	polyphenols	polar lipids	apolar lipids	organic acids	sugars	No class (YPred < 0)
2	phenols	64	84.38%	54	0	0	0	10	0	0	0	0	0
3	amines	31	61.29%	2	19	5	2	0	0	0	3	0	0
4	aromatic amines	47	57.45%	8	0	27	1	8	0	0	2	1	0
5	aminoacids	38	39.47%	0	7	10	15	0	0	0	1	5	0
6	polyphenols	125	96%	5	0	0	0	120	0	0	0	0	0
7	polar lipids	40	32.5%	5	0	0	2	0	13	20	0	0	0
8	apolar lipids	40	97.5%	0	0	0	0	0	1	39	0	0	0
9	organic acids	29	82.76%	3	1	0	0	0	0	0	24	1	0
10	sugars	32	93.75%	0	0	0	0	0	0	0	2	30	0
11	No class	0		0	0	0	0	0	0	0	0	0	0
12	Total	446	76.46%	77	27	42	20	138	14	59	32	37	0

Table 1: The misclassification table of the compound classes. The number of compounds assigned to the correct class is underlined in green, the ones assigned to the wrong class in yellow.

In the misclassification table, five of the chemical classes have prediction accuracy above 80%, 2 above 55% and 2 below 40%. The overall accuracy is 76%, that is, so far, not so bad. On the other hand, I must consider that the suggested thresholds for assignment to a class (0.3) and to unique assignment to a class (0.5) have not been used. This means that this data is over-estimated. This was so far the best achievement in classification that I have obtained from the PLS-da. I tried to classify the metabolites according to different parameters (biosynthetic pathway, or using wider or narrower chemical classes) but for many reasons not discussed here, it was impossible to me to go beyond this value. The idea I got from the model is that the classification of the metabolites in different classes is too weak; many metabolites are structurally similar to each other, and their separation in different chemical classes makes no sense. As example, I report the structure of Amygdalin: according to the rules of the classification stated above, amygdalin falls in the aromatic amine group. Nevertheless, from its structure, it is obvious that it is very similar to other phenolic compounds, with the only difference of having an immino group (image 11).

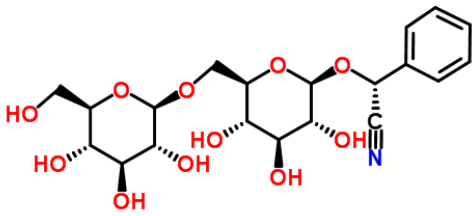


Image 11: Amygdalin, An Immino-phenol diglucoside commonly found in the seeds of many fruits. The structure here reported is from the ChemSpider repository: www.chemspider.com

Seems legit that for this kind of compounds the proposed classification strategy is misleading, and that the classification approach should be inclusive (one metabolite could be part of multiple classes), but this would not be so informative from my point of view. One of the explored solutions would be to create narrower classifications, but then the problem of number of replicates arises. On the other hand, the use of wider classification is useless. Furthermore, the main idea of this method was to improve the identification of the metabolites and not to have a dry classification, which could not really help on the identification. So this strategy was abandoned, and a wiser approach was used: the regression approach.

6.2.2.2 Regression approach

In the regression approach, the goal of the model is not the direct classification of the metabolites, but the creation of a regression model between the X matrix and the subunits that form the structure of metabolites. Indeed every chemical structure is composed of multiple substructures, like phenolic groups, aliphatic chain, and acidic substituents and so on. In this approach, the subunits are the Y matrix of the model. To create a regression model, a Partial Least Squares regression (PLSr) is performed using the "pls" R package from Mevik et al. (2013). The first step is to create a Y matrix of the compounds dataset according to their chemical subunits. Thirteen distinct Y variables have been chosen according to the variability present in our in-house dataset, here listed:

1) Polymeric structure (pol.str.) → the presence in the structure of two clear distinguishable subunits
2) Aliphatic chain (ali.cha.) → an aliphatic chain with a length included between 4 to 12 CH_2 units
3) Long aliphatic chain (l.ali.cha.) → an aliphatic chain longer than 12 CH_2 subunits
4) Aromatic ring (aro.gro.) → the presence of resonating rings of whichever nature
5) Homo-cycles (hom.cyc.) → the presence of phenolic rings in the structure
6) Hetero-cycles (het.cyc.) → the presence of resonating rings containing different atoms than Carbon and Hydrogen
7) Presence of Nitrogen (pre.nit.) → the presence of Nitrogen in the structure
8) Number of Nitrogens (n.N.) → the number of Nitrogens in the structure
9) Presence of Sulfur (pre.Sul.) → the presence of Sulfurs in the structure
10) Presence of Phosphorous (pre.Pho.) → the presence of Phosphorous in the structure
11) Glycosidic moieties (gluc.) → the number of glycosides present in the structure
12) Acidic group (Ac.gro.) → the presence of an acidic group in the structure
13) Aliphatic substituent (Sho.cha.) → a side aliphatic substituent of length minor than 4 CH_2

Starting from these 13 parameters, I created a Y matrix for all the metabolites manually checking their structure from the repository ChemSpider (www.chemspider.com). An automatic attempt of extraction of the substructures from the SMILES code has been also attempted. The results of this approach are the goal of this chapter and are described deeper in the next section.

6.3 Results: validation, error influence and external validation.

6.3.1 Data pre-treatment

The PLSr statistical test was used in this analysis. A straightforward description of necessary steps to achieve a reliable and robust data description is given in the paper of Wold et al. (2001); the 10 steps descripted have been used as reference in my work. One important point is to exclude the outliers in the analysis, which may weak the statistical power of the test. To mine the outliers, a useful tool is the Robust PCA analysis from the package "rrcov" (Todorov & Filzmoser, 2009), a PCA method that determines the distances of the observations according to the median, instead of the mean. After Robust PCA analysis, The compounds with a score value about 30 have been excluded (Image 12).

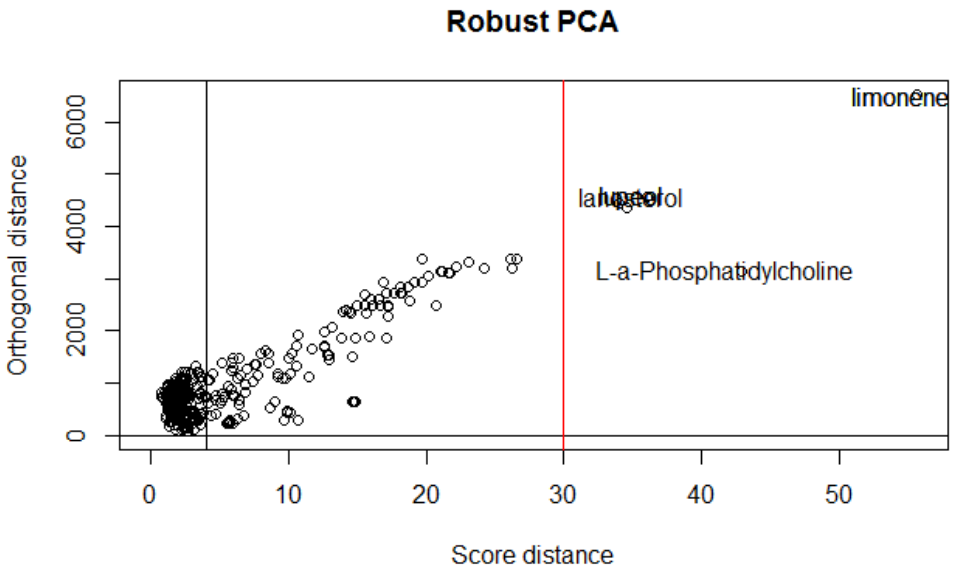


Image 12: Robust PCA of the whole dataset of standards, injected in my LC-MS instrument. The compounds above 30 in the score distance have been considered as outliers and excluded from the further analysis.

6.3.2 Method validation

PLS regression analysis have been performed using the X and Y matrices from the remaining compounds. The dataset of 439 metabolites have been fit in the model using the o-score-pls function of the "pls" package (the diverse algorithm have been tested and o-score-pls gave the best results). The method gave in automatic 9 different components; to avoid data over fitting, the RMSEP of every Y has been tested and I concluded that the exclusion of the last component, does not affect the accuracy of the predictions. The model run with the 8 components gave good explained variances for many Ys,

reported in table 2. The total variance of the X matrix explained by the model is 99.51%. The Y responses with values above 50% are considered well predicted. The "pls" package enforces these results performing Cross-validation on the dataset, and the cross-validated results were not different from these ones, indicating that the model is robust and more likely is not suffering of over-fitting. Nevertheless the 50% threshold is just a conventional level and not an absolute value; moreover cross-validation is a reliable tool, but it can be biased (Hall & Marron, 1987). So the dataset must be tested with a validation set, to evaluate better the absence of over fitting and the prediction properties of the model.

TRAINING: % variance explained after each component								
	1° comp	2° comp	3° comp	4° comp	5° comp	6° comp	7° comp	8° comp
X	30.50904	67.981	76.13	87.503	95.469	98.56	99.03	99.51
pol.str	52.0689	55.608	55.894	56.261	56.497	56.54	56.6	59.18
ali.cha	0.04208	66.886	69.906	70.029	70.032	74.19	74.41	74.67
l.ali.cha	2.7304	55.724	63.402	63.507	64.449	73.56	74.1	78.2
aro.gro	55.78775	64.322	75.949	78.686	79.528	85.08	89.84	90.61
Hom.cyc	58.74286	64.525	76.253	78.206	79.437	85.4	93.13	94.71
Het.cyc	2.51668	4.367	6.705	11.926	12.356	13.56	19.21	19.24
pre.nit	30.51252	31.669	31.872	64.045	65.416	65.65	65.68	65.84
n.N	9.88755	11.964	14.975	25.923	26.646	26.74	27.36	29.84
Pre.sul	0.73801	1.614	23.256	43.891	97.616	99.98	99.98	100
pre.Pho	0.08192	3.992	4.974	5.817	10.95	13.73	14.84	18.41
gluc	11.55042	19.335	30.113	30.357	42.645	43.49	50.75	66.07
Ac.gro	4.16389	4.7	7.788	7.869	8.363	15.81	15.82	17.64
Sho.cha	1.11194	19.627	19.722	20.642	21.033	24.95	24.97	25.14

Table 2: The explained variance after each component of the X matrix (first value) and of every Y response. In PLSr analysis, the threshold indicating good regression is 50%.

In the validation analysis, the dataset of compounds splits in two, a training set and a validation set. The validation set is composed of 30 random compounds from the whole set, and the remaining 409 are used as training set. In this test, training set is used to train the model, and the validation set Y responses are then predicted from the model and compared to the original values. The difference between the predicted values and the real ones gives an estimation of the model prediction power. A more appropriate validation test consists in the iteration of the random splitting per 1000 times, trying to predict the Y responses every time and measuring the average and the standard deviation of the percentage of correct predictions performed by the model.

The model built with the PLSr statistical test has been evaluated with the test described above. 1000 random validation sets have been created and they have been evaluated according to the prediction values for each Y response. As the predictions were not integer numbers, the values have been round to the closest integer. The predicted values were subtracted from the real values and averaged across the 1000 predictions; it was thus calculating the average of correct assignments across the 1000 iterative tests (parameter A). It showed good prediction properties (table 3).

	8 components	
Y matrix	average %	σ%
pol.str.	89.43	5.68
ali.cha.	95.85	3.56
l.ali.cha.	99.52	1.22
aro.gro.	86.31	6.07
Hom.cyc.	90.17	5.28
Het.cyc.	89.71	5.61
pre.nit.	91.65	4.97
n.N.	68.68	8.34
Pre.sul.	100.00	0.00
pre.Pho.	97.32	2.90
gluc.	82.22	6.87
Ac.gro.	68.79	8.02
Sho.cha.	80.50	6.82

Table 3: Percentage of the averaged correct values predicted by the model during 1000 iterative rounds. In the table also the standard deviation is reported.

The table 3 clearly shows that the there is a relationship between the X and Y matrices. Anyways, during the evaluation of a new model, it is necessary to consider the probability that the existing relationship is due to chance or is only apparent. To evaluate the performance of the model in this test, it is necessary to know how the model would perform when there is not relationship between X and Y matrices in the validation set. To break the supposed relationship between X and Y matrices, permutation test has been performed (Lindgreen et al. 1996). In the permutation test, the Y matrix is confused, scrambling the Y values of the variables. The result is that if relationship between X and Y matrices exists, this relationship disrupts after scrambling the variables. In the image 12, a comparison between the permuted and un-permuted data is shown.

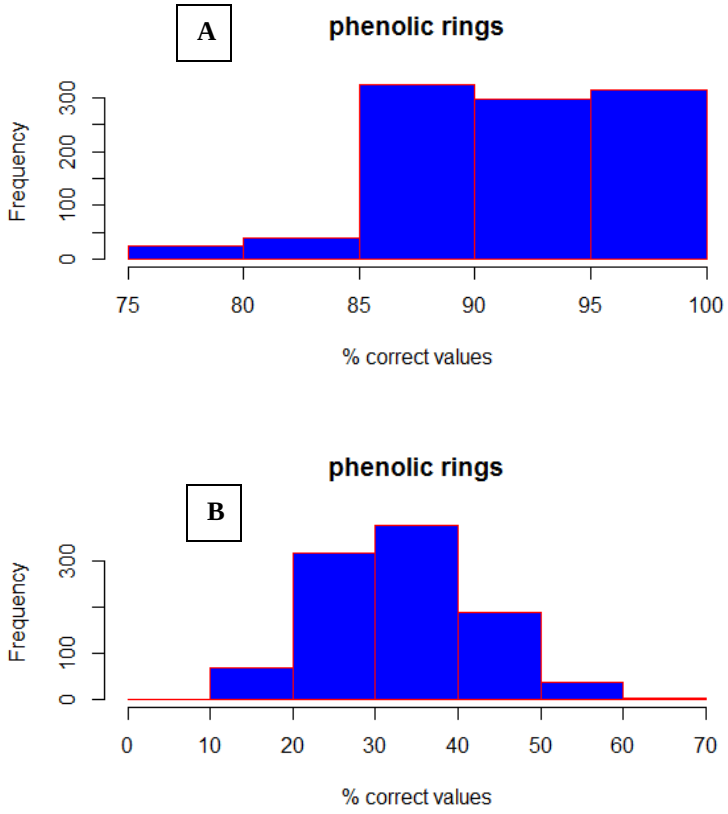


Image 12: A comparison between un-permuted (A) and permuted (B) percentage of correct values for the phenolic rings (homocycles). The distributions are reported as histogram of correct values, across 1000 times test replicates. The data shows clearly that the permuted data has a normal distribution and random accuracy, while the un-permuted data has not a normal distribution and is one-side tailed.

The different distributions have been tested with Mann-Whitney U test to see if the differences are statistically relevant or not. Results are shown in table 4. As shown in the table, two responses did not pass the test, the statistical test could not refuse the null hypothesis (distribution is the same for both data).

%	*un-permuted data*	*permuted data*
pol.str.	**89.10**	51.53
ali.cha.	**95.47**	71.19
l.ali.cha.	**99.62**	83.38
aro.gro.	88.03	32.51
Hom.cyc.	**92.45**	34.71
Het.cyc.	89.77	89.84
pre.nit.	**91.44**	65.59
n.N.	69.93	51.75
Pre.sul.	**100.00**	94.44
pre.Pho.	97.47	97.58
gluc.	83.32	58.76
Ac.gro.	68.58	49.59
Sho.cha.	80.50	69.33

Table 4: A comparison of the number of correct predictions between un-permuted and permuted data. The yellow labeled data indicates the responses that do not show any difference in the distribution between the permuted and un-permuted data.

Nonetheless, the values obtained from the permuted test did not correspond to the random prediction of the values of the Y matrix. For example, "l.ali.cha." has two possible values: "0" (absence) and "1" (presence), therefore, the expected random assignment would have an accuracy of about 50%. In my case, the permuted accuracy was in average 83.38% (table 4). Due to the high number of 0 in the values of the responses of "l.ali.cha", I suspected that the model was assigning always "0" and it had false predictive properties. Therefore, the estimation reported in table 4 based on the parameter A is necessary to demonstrate that there is prediction power in the model, but it is non-sufficient. It needs other parameters to understand better the predictive power of the model.

To overcome this limit, I performed another test, calculating the number of non-zero response values predicted correctly from the model (parameter B). Results are shown in table 5. As shown, the ability of the model to predict the non-zero values is a lot higher than in the permuted data, indicating a good prediction behavior. As observed in the previous test, the model is not able to predict the heterocycles and the presence of Phosphorous. Nonetheless, there are another two parameters scarcely predicted: the number of Nitrogens and the short aliphatic substituent, which have a prediction value below 50% (it is more effective to give random chances than to use the model in this case).

%	un-permuted data	permuted data
pol.str.	78.18	38.40
ali.cha.	78.77	15.77
l.ali.cha.	**97.73**	10.26
aro.gro.	85.51	23.69
Hom.cyc.	**95.92**	23.67
Het.cyc.	0.44	0.13
pre.nit.	69.27	19.28
n.N.	40.57	20.00
Pre.sul.	**100.00**	4.90
pre.Pho.	0.00	0.00
gluc.	66.54	21.98
Ac.gro.	63.15	40.67
Sho.cha.	37.16	13.67

Table 5: A comparison of the number of correct predictions of the non-zero responses between un-permuted and permuted data. The yellow labeled data indicates the responses that do not show any difference in the distribution between the permuted and un-permuted data. The bold numbers indicate very good predictive power of the model.

On the other hand, I would like to underline the extremely good performance on the prediction of the presence of long aliphatic chains, the phenolic rings (homocycles) and the presence of Sulfur, indicating that the X predictions strongly relate with these Y responses. In particular, the phenolic rings

is very interesting; in the model there are up to 9 possible values that the model can predict, and out of these 9 answers it gets the correct one in the 95% of the cases.

In table 6, I performed the same test, calculating the average of wrong prediction of zero values (parameter C). In this test, higher the value, lower is the prediction power of the zero Y responses. As shown, the model has good capacity to predict the absence of "pol.str.", "ali.cha.", "l.ali.cha.", "pre.nit." and "Pre.sul.". The bad predictions for Het.cyc and Pre.Pho confirmed also in this test.

%	un-permuted data	permuted data
pol.str.	**0.67**	38.24
ali.cha.	**0.21**	15.56
l.ali.cha.	**0.26**	9.24
aro.gro.	8.62	57.45
Hom.cyc.	11.23	55.60
Het.cyc.	0.06	0.08
pre.nit.	**0.55**	18.45
n.N.	19.63	38.01
Pre.sul.	**0.00**	2.92
pre.Pho.	0.00	0.00
gluc.	10.81	28.52
Ac.gro.	27.16	44.46
Sho.cha.	5.36	12.70

Table 6: A comparison of the number of wrong predictions of the zero responses between un-permuted and permuted data. The yellow labeled data indicates the responses that do not show any difference in the distribution between the permuted and un-permuted data. The bold numbers indicate very good predictive power of the model.

Next step in the analysis was to consider the error in the measurements that the LC-MS instruments make during data acquisition. The error in the measurements is multifactorial, and instrument-specific. There are some general considerations on the error that needs to be taken in account, according to the type of instrument used. In my experiments, I used SYNAPT G1 mass spectrometer (WATERS, Manchester, UK). It is a Q-TOF with a maximum resolution of 17.500 measured on the (M + 6H)^6+ isotope cluster from bovine insulin (m/z 956) in negative and maximum resolution of 17.500 measured on the (M - 4H)^4- isotope cluster from bovine insulin (m/z 1431) in positive. The vendor accuracy of above assures an error less than 2 ppm in mass accuracy, in absence of 1) interference and 2) enough intensity.

1) Interfering analytes can be whichever analyte has an ion of m/z close enough to the ion of interest under analysis (Thurman & Ferrer, 2010). "Close enough" depends from the resolution of the instrument. Due to the untargeted nature of the untargeted analysis, in theory

interference can be everywhere in the chromatograms and MS spectra, it is not possible to avoid it, and very difficult to recognize.

2) Shahaf et al. (2013), reported the construction of a mass accuracy error surface according to ion intensity, based on the same instrument here used. In their work, they report that in most cases the error is around 5 ppm, has a 95% confidence interval below 30 ppm, but can reach very high values (100 ppm) when very low intensity ions are measured. Even in new instruments, including last version of Orbitrap, errors over 10 ppm are reported when the intensity is low (Knolhoff et al. 2014).

Isotopic intensities measurements also suffer of error in their estimation. The error is instrument-specific. Generally TOF instruments are reported to have a better estimation of the intensities (Kind & Fiehn, 2006) with an error around 2%, while Orbitrap can have errors above 5%. Nevertheless, the error in the real data can be bigger (Knolhoff et al. 2014).

In this study, I evaluated the effect of both mass accuracy error and Carbon content prediction error on the regression model. Results displayed below. First, the mass measurement error has been evaluated. The mass values of the validation set have been added of ± 5, 10, 30, 50, 100 ppm of mass error, and a curve has been built using the values of correct Y predictions. In this test, the model showed three main behaviors: 1) Mass error dependent predictions. 2) Mass error independent predictions. 3) bad predictions. Examples of the three behaviors are reported in the Image 13.

The mass error dependent predictions showed a very interesting result. The prediction power of the model was not dropping with ten ppm error, and it was still acceptable with 30 ppm error. This is probably the main finding of this study: the combined use of multiple predictors can overcome the error in single or few of them. In this case, the mass error was affecting the predictors 4 (mass defect), 5 (relative mass defect) and 8 (residual relative mass defect). Despite of three out of nine predictors had wrong values, the predictions were correct for up to 90% of the cases when ppm error was 10 ppm. It is redundant to say that with 10 ppm error, it is already very difficult (if not impossible), calculate the chemical formula (Kind & Fiehn, 2007).

If the first and the third group were clearly expected, the presence of a group of responses not affected by the mass error was unexpected. The presence of this group indicates that some Y responses are orthogonal to some X predictors. This important finding will be demonstrated in the next section.

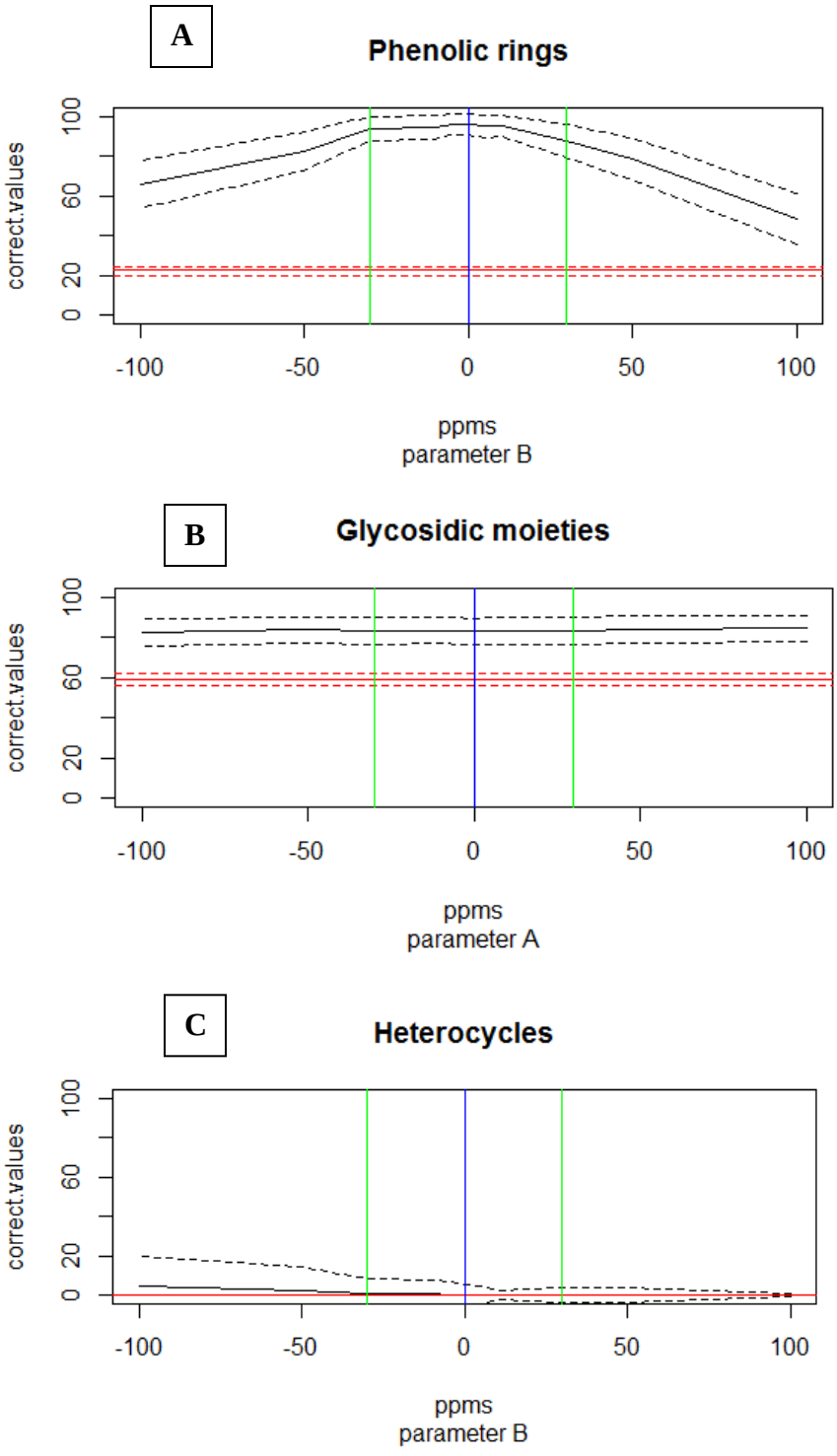


Image 13: The curves represent the amount of correctly predicted non-zero values across 1000 iterations of the test (parameter A & B), with error value ranging from -100 ppm to 100 ppm. The dashed line indicates the standard deviation of the predictions, the green lines indicate the 30 ppms limits. The red lines are the average, minimum and maximum values obtained from the permutation tests.

The phenolic.rings (A) is a good example of "mass error dependent response", with the amount of predictions dropping with mass errors superior to 30 ppms.

The aliphatic chains (B) is an example of "mass error independent response", having the prediction value slightly varying across the mass errors.

The heterocycles (C), and the presence of Phosphorous showed to be unpredictable from the model.

Polymeric structure, presence of Nitrogen, presence of Sulfur, Glycosidic moieties, Acidic groups and Short chain were mass error independent responses. Aromatic rings, aliphatic chain, long aliphatic chain and Phenolic rings were mass error dependent responses, and the remaining ones (heterocycles, number of Nitrogens, and presence of Phosphorous) were bad predicted.

The same test has been performed on the Carbon estimation errors *Cees* (the calculation of the Carbon content will be described in the next section). It showed similar results to the mass measurement error. The accuracy in the determination of the number of Carbons is necessary to predict many responses. Some of the responses showed a peculiar trend, shown in the Image 14.

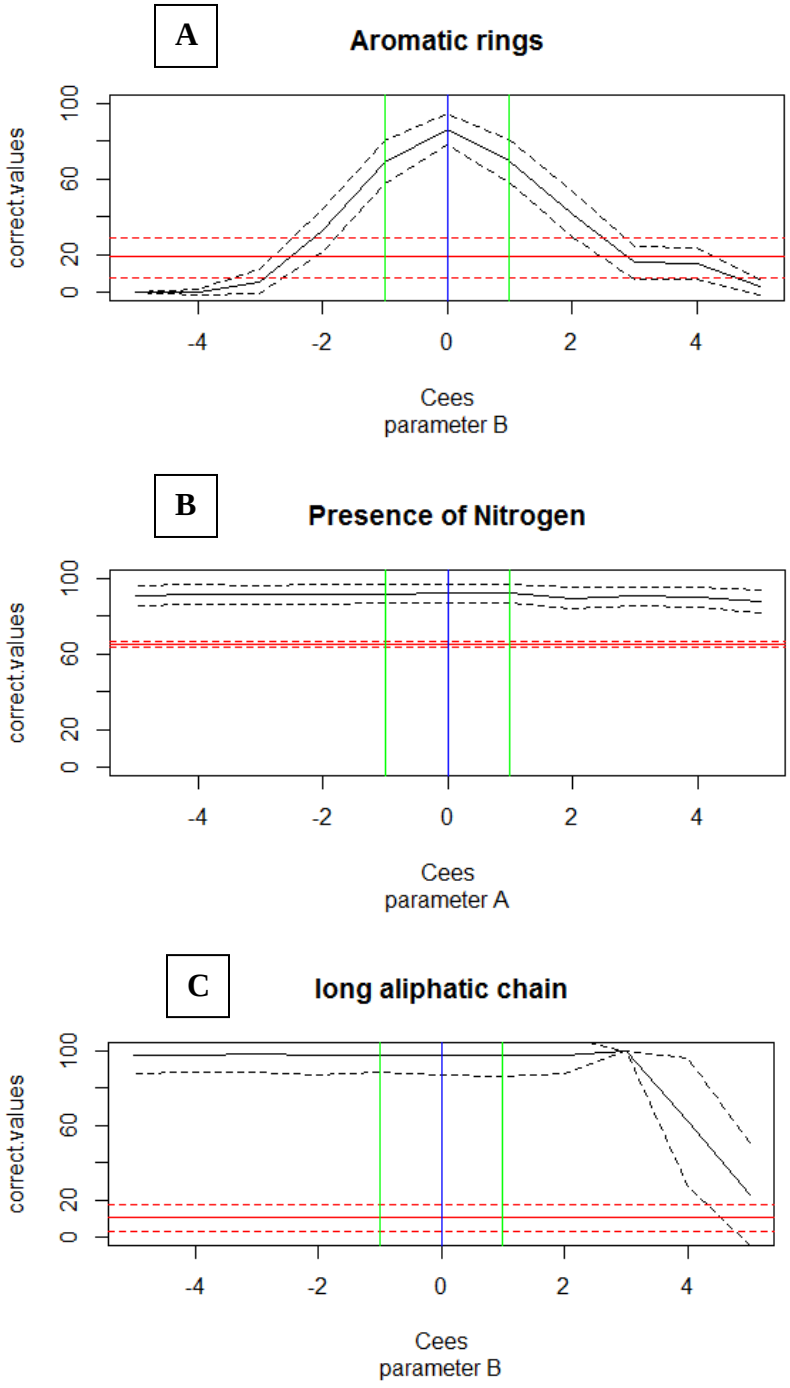


Image 14: The curves represent the amount of correctly predicted non-zero values across 1000 iterations of the test (parameter B), with error value ranging from -5 to + 5 Cees (Carbon Estimation Errors). The dashed line indicates the standard deviation of the predictions, the green lines indicate the 1 Carbon limits. The red lines are the average, minimum and maximum values obtained from the permutation tests.

The aromatic rings (A) is a good example of Carbon estimation error dependent response, with the error increasing in both sides of the error surface.

The presence of Nitrogen (B) is a good example of Carbon estimation error independent response, with the predictions unaffected or slightly affected by the error.

The long aliphatic chain (C) is an example of high positive error depndent response. In this class the estimation was dropping with values above 3 Carbons.

The aromatic rings, phenolic rings, Glycosidic moieties, and the short chains were Carbon estimation error dependent responses. The presence of Nitrogen, short chain and the presence of Sulfur were carbon estimation error independent responses. The heterocycles and the presence of Phosphorous were bad predicted response. In this test, there were also responses that showed a peculiar trend; their predictability was dropping only in one side of the Carbon estimation error. The Polymeric structure, aliphatic chain and long aliphatic chain were dropping with Carbon estimation error above 3, probably because the values were overtaking the maximum number of carbons (calculated as Nominal mass/12), and were creating impossible combinations. The number of Nitrogens and the acidic group were dropping with Carbon underestimation of two units.

RT error has been evaluated as well. Errors in the predictions have been found with values above 10 minutes. In real chromatographic conditions, 10 minutes shift is impossible, so I concluded that the error in RT shift is not a limiting factor.

In conclusion, I found that errors below 10 ppm in the mass measurement, and below ± 1.5 Carbons in the carbon estimation allow the model to have good prediction properties and error in estimation of the substructures below 20% of the cases (apart for the ones bad predicted). Last in this section, an experimental condition like estimation has been performed, using as errors the normal distribution around the correct values of 10 ppm and 1 carbon error as standard deviation. Results are displayed in table 7.

%	Parameter A	permuted data	Parameter B	permuted data	Parameter C	permuted data
pol.str	88.67	51.53	77.78	38.40	1.22	38.24
ali.cha	95.53	71.19	80.38	15.77	0.78	15.56
l.ali.cha	98.85	83.38	97.35	10.26	1.01	9.24
aro.gro	76.62	32.51	71.49	23.69	16.21	57.45
Hom.cyc	81.21	34.71	76.61	23.67	13.99	55.6
Het.cyc	89.91	89.84	4.08	0.13	0.58	0.08
pre.nit	91.38	65.59	68.82	19.28	0.72	18.45
n.N	68.35	51.75	40.80	20.00	21.85	38.01
Pre.sul	100.00	94.44	100.00	4.90	0.00	2.92
pre.Pho	97.44	97.58	0.00	0.00	0.07	0
gluc	82.38	58.76	63.78	21.98	11.00	28.52
Ac.gro	65.43	49.59	59.66	40.67	30.05	44.46
Sho.cha	80.24	69.33	37.50	13.67	5.81	12.7

Table 7: The results of the test with normal distributed error in mass measurement and Carbon content estimation.

The table 7 shows that the model suffers a slight loss of predictability when the normal distributed error is applied. The main parameters suffering from the error are the number of aromatic rings, the number of phenolic rings and the number of glycosides. The first two parameters have both 9 possibilities (from 0 to 8), so this means that when the error is applied, an increased number of compounds are predicted with the wrong amount of rings. The third one has 3 different possibilities. On the other hand, I also calculated the average of difference between the predicted number of these responses and the real number, and I found that the average error is 1.10, indicating that most of the times, the difference between predicted and the real number is only 1.

The remaining predicted responses showed non-significant differences with the ones from the previous tests, because as shown earlier they are unaffected by the errors, especially if the error is small. The orthogonality between these X predictors and Y responses is shown in the next sections.

6.3.3 Carbon content prediction

One of the main parameter of this model is the calculation of the number of Carbons present in the chemical formula of the compounds. Due to its peculiar isotopic distribution, Carbon results to be one of the easiest element to be calculated. In nature, there are only two isotopes of carbon stable: $Carbon^{12}$ and $Carbon^{13}$. The distribution of the isotopes is 98.9% for the first one and 1.10% for the second one. To give an example and show how precise this relationship is, in image 15 is shown the theoretical linear regression between the number of carbons and the ratio between the first and the second isotope of the whole compound set used in this experiment. As you can see, the regression is linear, with a R^2 of 0.9969 and a standard error of 0.4851.

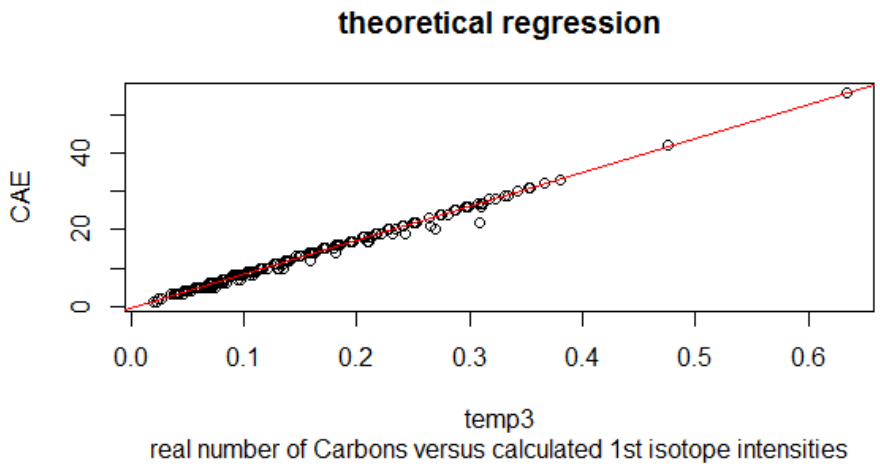


Image 15: The theoretical regression between the number of carbons (Y axis) and the ratio between the two carbon isotopes (X axis). The R^2 = 0.9969 and the standard error is only 0.4851.

In this model, I needed a method to calculate quickly and reliably the number of Carbons. To extract the isotopic intensities from the peaktable, I used the script developed by Dr. Jan Stanstrup based on the intensity-weighted means, described by Stanstrup et al. (2013). This method is very reliable in the data extraction and has been demonstrate to have an average error of 2.34% in comparison to the theoretical ratio. As I needed only the number of Carbons, I used the linear model showed in Image 15 to calculate the number of Carbons, using the values from the intensity weighted mean method.

6.3.4 Orthogonal factors of the model

As shown in section 5.3.3, some of the Y responses were unaffected from the error in the mass measurement and carbon content. To confirm this finding, the model has been rebuild excluding the values from the mass accuracy, i.e. mass defect, relative mass defect and residual relative mass defect (md, RMD, rRMD) and the values from the Carbon content, i.e. number of Carbons, percentage of Carbons and residual relative mass defect (nofC, pC, rRMD). The values from the training of these models have been compared to the original one, to see which Y responses are affected by these X predictors. Results are shown in table 8.

	original data	*no mass accuracy*	*no carbon content*
Y responses	*8 comps*	*6 comps*	*6 comps*
pol.str	59.18	53.96	58.01
ali.cha	74.67	48.32	71.29
l.ali.cha	78.20	46.02	75.75
aro.gro	90.61	59.59	67.75
Hom.cyc	94.71	62.60	65.94
Het.cyc	19.24	14.90	14.10
pre.nit	65.84	62.82	65.77
n.N	29.84	25.97	25.63
Pre.sul	100.00	100.00	100.00
pre.Pho	18.41	20.27	8.17
gluc	66.07	60.89	41.26
Ac.gro	17.64	14.62	11.50
Sho.cha	25.14	19.46	23.82

Table 8: In the table, the Y variance explained from the X are compared between the original model (2nd row), the model with **md** information (3rd row) and the model without Carbon content informations (4th row). The red color underlines the values that drop after the exclusion of some x predictors from the model. As you may notice, some of the Y responses are not affected by the absence of some X predictors, indicating that this information is orthogonal to the Y responses

The table shows that some of the responses are almost unaffected from the lack of some of the X predictors. This means that their information is orthogonal or almost orthogonal to the responses. On

the other hand, the aromatic rings and the phenolic rings are very affected from the lack of some predictors.

6.3.5 Predictability of the test set.

In model building, the last step is to evaluate the performance of the model on an independent test set (Szymanska et al. 2012), a set not used for the creation of the model. In this project, my idea has been to obtain a test set from real data (grape samples), instead of in-silico data. In facts, I wanted to evaluate also if the model is capable to predict structures in real cases and how big is the error of prediction in the real samples.

In the experiment described in chapter 6, I identified the unknown compounds using their MS/MS spectra, their sum formula and the results from this model. Nevertheless, 19 of the metabolites identified were already reported to be present in grape, so their presence was expected and their identification at MSI level 2 was achieved simply comparing their spectra to the one stored in external databases. These 19 metabolites have been used to see how the model here described predicts their substructures (Y responses). Parameter A has been used in this analysis. Results are in table 9.

The table shows that in most of the cases the Y predicted corresponds to the Y real responses. “Polymeric structure”, “long aliphatic chain”, “presence of Nitrogen”, “presence of Sulfur” and “presence of Phosphorous” did not suffer of any error (the reason why presence of Phosphorous had no error is because there were not compounds with phosphorous in their formula, but we know that this parameter cannot predict the “1” in its response). “Aliphatic chain” had an error of 5.26%, “Aromatic group” of 26.5% and “Phenolic rings” of 15.76%. “Glycosidic moieties” and “Acidic Group” had an error of 21%, while “Short chain” had an error of only 5.26%. “Number of Nitrogen” had a 36% of error, and showed to be completely unreliable. It is interesting to notice that, apart from “number of Nitrogen”, all the other ones showed an error of estimation of only 1. This is a good result for the responses that have multiple choices (“Aromatic rings”, “Phenolic rings” and “glycosidic moieties”), indicating that even if the number is not correct, it is not misleading completely. It is interesting to notice that some of the parameters have a similar value to the normal distributed test (table 7).

The results here showed are not statistically relevant; the number of compounds under analysis is too low. It is intended to give an idea to the performance of the model on real data. Further tests are required to understand how far can the model go, which metabolites can be predicted and if there is any way to improve the model.

	pol.str	ali.cha	l.ali.cha	aro.gro	Hom.cyc	Het.cyc	pre.nit	n.N	Pre.sul	pre.Pho	gluc	Ac.gro	Sho.cha
laricitrin 3 -glucoside	0	0	0	0	0	0	0	0	0	0	0	0	0
phenyl-alanine	0	0	0	0	0	0	0	0	0	0	0	**1**	0
Caftaric acid_glutathione	0	0	0	0	0	0	0	0	0	0	**1**	**1**	0
malvidin-3,5 diglucoside	0	0	0	0	0	0	0	0	0	0	0	0	0
eriodictyol-7-glucoside	0	0	0	0	0	0	0	0	0	0	0	0	0
terpenol-pentosyl-glucoside	0	0	0	0	0	0	0	2	0	0	**1**	0	0
geranic acid-rhamnosyl-glucoside	0	**1**	0	**1**	0	0	0	0	0	0	**1**	0	0
ampelopsin D+quadrangolarin A	0	0	0	**1**	**1**	0	0	0	0	0	0	0	0
quercetin-rhamnoside	0	0	0	0	0	0	0	0	0	0	0	0	0
quercetin-glucuronide	0	0	0	**1**	**1**	0	0	2	0	0	0	**1**	0
terpendiol-rhamnosyl-glucoside	0	0	0	0	0	0	0	0	0	0	0	**1**	0
delphinidin-arabinoside	0	0	0	1	0	0	0	0	0	0	0	0	0
cyanidin-arabinoside	0	0	0	0	0	0	0	1	0	0	0	0	0
cyanidin-p-coumaroyl-glucoside	0	0	0	0	0	0	0	0	0	0	0	0	0
malvidin-diglucoside-acetate	0	0	0	0	0	0	0	0	0	0	0	0	**1**
p-coumaroyl-peonidin-diglucoside	0	0	0	**1**	**1**	0	0	2	0	0	**1**	0	0
myricetin	0	0	0	0	0	0	0	0	0	0	0	0	0
dihydrosyringetin-glucoside	0	0	0	0	0	0	0	0	0	0	0	0	0
catechin-glucoside	0	0	0	0	0	0	0	0	0	0	0	0	0
Percentage of error	**0.00%**	**5.26%**	**0.00%**	**26.32%**	**15.79%**	**0.00%**	**0.00%**	**36%**	**0.00%**	**0.00%**	**21.05%**	**21.05%**	**5.26%**

Table 8: The predicted Y for these nineteen compounds has been subtracted from the real Y responses and the results are summarized in this table. At the bottom, the percentage of wrong assignments. It is interesting to notice that the wrong assignment are always 1 and do not go beyond this value. Exception can be notice for n.N, which is a bad predicted parameter. The MS/MS spectra are shown in the supplementary table 4.

6.4 Conclusions and future outlooks

In this "proof of principles" study, I demonstrated that the combined use of multiple parameters in a regression model, built on compound database, is able to predict some substructures of the molecules under analysis. The regression model, built using Partial Least Squares regression was able to assign in most of the cases the right number of substructures, contributing on the classification of the compounds. Moreover, I demonstrated that the use of multiple parameter together in multivariate regression improve the model robustness, allowing the model to be almost unaffected in mass measurement error of ±10 ppm and isotopic intensities ratio errors of ±1.5 estimated Carbons. This result is not achievable with the standard measurement procedure described in the introduction (section 5.1).

In the study shown in chapter 6, I used this method manually with the MS/MS spectra and the formula calculated by the Rdisop package (Bockler & Liptak, 2007), to assign structures to the unknown biomarkers. It was very helpful to me; whenever I did not have a reliable candidate for the structure of one unknown biomarker, and the chemical formula from Rdisop was misleading, the combined use of the model and the MS/MS spectrum allowed me to sketch a putative structure and compare the in-silico fragmentation of such structure with the one suggested by the Chemspider database in MetFrag. I had only 50% of the biomarkers in my standard database, but I was able to give a reliable identification to 90% of them.

On the other hand, there are multiple limits in this method, and in the future, the development needs to solve the following problems:

1) The Y matrix has been manually built. This is not the ideal case, and it should be built in an automatic manner. One R package able to detect substructures in SDF files exists, "fmcsR" (Wang et al. 2014), which works in couple with "ChemmineR" (Cao et al. 2008). They have been tested. It is not able to detect all the substructures used in this work, it uses exclusively SDF files, while databases often furnish only Smiles and InChi codes and their translation to SDF is not straightforward.
2) I used the internal FEM database of compounds. The ideal case would be to use databases richer of compounds, to enforce the statistical power of the model. I tried to use external databases, but I had only access to Plantcyc grape compounds database (www.plantcyc.org), which is not ideal for this test, because it also contains proteins, metal ions, some formulas

are missing, some smiles are missing (some are wrong), therefore I did not achieve any result from it.

3) As stated above, the results from the model can be used manually coupling the outcome with the MS/MS spectra and any other information. This is not ideal. Any information should be integrated in the model, including MS/MS spectra. Therefore, the ideal would be to integrate all the informations possible in the model, to enforce its predictability. For example, in MS/MS spectra Phosphate, glycosides, and acidic group give specific fragments, which might enforce the prediction of these substructures. The UV spectra could be also very interesting to integrate in the model. The perfect database to use would be Massbank in this case (www.massbank.jp), where MS/MS spectra are stored.

References chapter 6

1. *Branden V. K., & Hubert, M. (2005). Robust classification in high dimensions based on the SIMCA Method. Chemometrics and Intelligent Laboratory Systems, 79, 10–21. doi:10.1016/j.chemolab.2005.03.002*
2. *Bjørn-Helge Mevik, Ron Wehrens and Kristian Hovde Liland (2013). pls: Partial Least Squares and Principal Component regression. R package version 2.4-3. http://CRAN.R-project.org/package=pls*
3. *Böcker, S., & Liptak, Z. (2007). Algorithmica A Fast and Simple Algorithm for the Money Changing Problem 1, 413–432.*
4. *Boswell, P. G., Schellenberg, J. R., Carr, P. W., Cohen, J. D., & Hegeman, A. D. (2011). A study on retention "projection" as a supplementary means for compound identification by liquid chromatography-mass spectrometry capable of predicting retention with different gradients, flow rates, and instruments. Journal of Chromatography. A, 1218(38), 6732–41. doi:10.1016/j.chroma.2011.07.105*
5. *Boswell, P. G., Schellenberg, J. R., Carr, P. W., Cohen, J. D., & Hegeman, A. D. (2011). Easy and accurate high-performance liquid chromatography retention prediction with different gradients, flow rates, and instruments by back-calculation of gradient and flow rate profiles. Journal of Chromatography. A, 1218(38), 6742–9. doi:10.1016/j.chroma.2011.07.070*
6. *Bylesio, M., Rantalainen, M., Cloarec, O., Nicholson, J. K., Holmes, E., & Trygg, J. (2007). OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. Journal of Chemometrics, (February), 398–405. doi:10.1002/cem*
7. *Cao, Y, Charisi, A, Cheng, L C, Jiang, T, Girke, T (2008) ChemmineR: a compound mining framework for R. Bioinformatics, 24: 1733-1734. URL: http://www.hubmed.org/display.cgi?uids=18596077.*
8. *Creek, D. J., Jankevics, A., Breitling, R., Watson, D. G., Barrett, M. P., & Burgess, K. E. V. (2011). Identification by Retention Time Prediction. Analytical Chemistry, (83), 8703–8710.*
9. *Gerlich, M., & Neumann, S. (2013). MetFusion: integration of compound identification strategies. Journal of Mass Spectrometry : JMS, 48(3), 291–8. doi:10.1002/jms.3123*

10. *Hall, P., & Marron, J. S. (1987). Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. Probability Theory and Related Fields, 74, 567–581. doi:10.1007/BF00363516*
11. *Kind, T., & Fiehn, O. (2006). Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. BMC Bioinformatics, 7, 234. doi:10.1186/1471-2105-7-234*
12. *Knolhoff, A. M., Callahan, J. H., & Croley, T. R. (2014). Mass accuracy and isotopic abundance measurements for HR-MS instrumentation: capabilities for non-targeted analyses. Journal of the American Society for Mass Spectrometry, 25(7), 1285–94. doi:10.1007/s13361-014-0880-5*
13. *Kramer R. W., E.B. Kujawinski, X. Zang, K.B. Green-Church, B. Jones, M.A. Freitas, P.G. Hatcher. Studies of the structure of humic substances by electrospray ionization coupled to a quadrupole-time of flight (QQ-TOF) mass spectrometer. Spec. Publ. - R. Soc. Chem. 2001, 273, 95.*
14. *L. Sta°hle, S. Wold, (1987). Partial least squares analysis with cross-validation for the two-class problem: a Monte Carlo study, J. Chemom. 1. 185–196.*
15. *Lindgren, F., Hansen, B., & Karcher, W. (1996). Model validation by permutation tests: Applications to variable selection. Journal of Chemometrics, 10, 521–532.*
16. *Rasche, F., Svatos, A.,Maddula, R.K., Böttcher, C., and Böcker, S. (2010). Computing fragmentation trees from tandem mass spectrometry data. Anal. Chem. 83, 1243–1251. doi:10.1021/ac101825k*
17. *Rogers, S., Scheltema, R. a., Girolami, M., & Breitling, R. (2009). Probabilistic assignment of formulas to mass peaks in metabolomics experiments. Bioinformatics, 25(4), 512–518. doi:10.1093/bioinformatics/btn642*
18. *Shahaf, N., Franceschi, P., Arapitsas, P., Rogachev, I., Vrhovsek, U., & Wehrens, R. (2013). Constructing a mass measurement error surface to improve automatic annotations in liquid chromatography/mass spectrometry based metabolomics. Rapid Communications in Mass Spectrometry, 27(21), 2425–2431. doi:10.1002/rcm.6705*
19. *Silva, R. R., Jourdan, F., Salvanha, D. M., Letisse, F., Jamin, E. L., Guidetti-Gonzalez, S., … Vêncio, R. Z. N. (2014). ProbMetab: An R package for Bayesian probabilistic annotation of LC-MS-based metabolomics. Bioinformatics, 30(9), 1336–1337. doi:10.1093/bioinformatics/btu019*

20. *Sleno, L. (2012). The use of mass defect in modern mass spectrometry. Journal of Mass Spectrometry, 47(2), 226–236. doi:10.1002/jms.2953*
21. *Stanstrup, J., Gerlich, M., Dragsted, L. O., & Neumann, S. (2013). Metabolite profiling and beyond: approaches for the rapid processing and annotation of human blood serum mass spectrometry data. Analytical and Bioanalytical Chemistry, 405(15), 5037–48. doi:10.1007/s00216-013-6954-6*
22. *Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. a., … Viant, M. R. (2007). Proposed minimum reporting standards for chemical analysis. Metabolomics, 3(3), 211–221.*
23. *Szymanska, E., Saccenti, E., Smilde, A. K., & Westerhuis, J. (2012). Double-check : validation of diagnostic statistics for PLS-DA models in metabolomics studies. Metabolomics, 8, 3–16. doi:10.1007/s11306-011-0330-3*
24. *Thurman, E. M., & Ferrer, I. (2010). The isotopic mass defect: A tool for limiting molecular formulas by accurate mass. Analytical and Bioanalytical Chemistry, 397(7), 2807–2816. doi:10.1007/s00216-010-3562-6*
25. *Todorov V., Filzmoser P. (2009). An Object-Oriented Framework for Robust Multivariate Analysis. Journal of Statistical Software, 32(3), 1-47. URL http://www.jstatsoft.org/v32/i03/.*
26. *Van Krevelen D. (1950). Graphical-statystical method for the study of structure and reaction process of coal. Fuel, 29, 269.*
27. *Vessecchi, R., Crotti, A. E. M., Guaratini, T., Colepicolo, P., & Galembeck, S. E. (2007). Radical Ion Generation Processes of Organic Compounds in Electrospray Ionization Mass Spectrometry. Mini-Reviews in Organic Chemistry, 4(1), 75–87.*
28. *Werner, E., Heilier, J., Ducruix, C., Ezan, E., Junot, C., & Tabet, J. (2008). Mass spectrometry for the identification of the discriminating signals from metabolomics: Current status and future trends. Journal of Chromatography. B, Analytical Technologies in the Biomedical and Life Sciences, 871(2), 143–163. doi:10.1016/j.jchromb.2008.07.004*
29. *Wehrens, R., Weingart, G., & Mattivi, F. (2014). metaMS: an open-source pipeline for GC-MS-based untargeted metabolomics. Journal of Chromatography. B, Analytical Technologies in the Biomedical and Life Sciences, 966, 109–16. doi:10.1016/j.jchromb.2014.02.051*

30. *Wold S. (1976). Pattern recognition by means of disjoint principal component models. Pattern recognition, 8, 127-139.*
31. *Wold, S., Kettaneh-Wold, N., & Skagerberg, B. (1989). Nonlinear PLS modeling. Chemometrics and Intelligent Laboratory Systems, 7, 53–65. doi:10.1016/0169-7439(89)80111-X*
32. *Wolf, S., Schmidt, S., Müller-Hannemann, M., & Neumann, S. (2010). In silico fragmentation for computer assisted identification of metabolite mass spectra. BMC Bioinformatics, 11, 148.*
33. *Wu, Z., Rodgers, R. P., & Marshall, A. G. (2004). Two- and three-dimensional van krevelen diagrams: a graphical analysis complementary to the kendrick mass plot for sorting elemental compositions of complex organic mixtures based on ultrahigh-resolution broadband fourier transform ion cyclotron resonance. Analytical Chemistry, 76(9), 2511–6. doi:10.1021/ac0355449*
34. *Yan Wang, Tyler Backman, Kevin Horan and Thomas Girke (2014). fmcsR: Mismatch Tolerant Maximum Common Substructure Searching. R package version 1.6.5. http://manuals.bioinformatics.ucr.edu/home/chemminer*
35. *Zhang, H., Zhang, D., & Ray, K. (2003). A software filter to remove interference ions from drug metabolites in accurate mass liquid chromatography/mass spectrometric analyses. Journal of Mass Spectrometry : JMS, 38(10), 1110–2. doi:10.1002/jms.521*

7. Untargeted comparative analysis of the metabolomes of *Vitis* vinifera and four American *Vitis* species reveals big differences in the accumulation of polyphenols and aroma precursors

The lack of natural selection for *Vitis vinifera* grapes makes them increasing susceptible to pests. Massive use of pesticides is therefore required for their cultivation, increasing environmental and economic costs. Hybrids obtained from repeated backcrossing of high quality *Vitis vinifera* genotypes with different species of the *Vitis* germplasm have been shown to be a viable solution in terms of introducing natural resistance characteristics to new cultivars suitable for the production of wines. However, lack of knowledge about the genomic and metabolomic resources available for each *Vitis* species makes the hybridization process very lengthy and not cost-effective.

In this proof of principle study, we analyzed the metabolome of some *Vitis* species berries, including *Vitis vinifera,* with the scope of identifying the metabolites that differentiate *vinifera* grapes from others. The results show that several metabolic differences exist and that some American *Vitis have* interesting characteristics, including some undesirable traits that should be taken into account in the design of breeding programs. The method suggested in this study could be considered in order to improve the design of new breeding programs, lowering the risk of retaining undesirable characteristics in the chemical phenotype of the offspring.

This work is part of the paper: Narduzzi L., Mattivi F. (2015) "Untargeted comparative analysis of the metabolomes of *Vitis* vinifera and four American *Vitis* species reveals big differences in the accumulation of polyphenols and aroma precursors" (In preparation)

7.1 Introduction:

Most of the grapes produced globally are *Vitis vinifera.* This species has its origin in the Near East, as the domesticated progeny of *Vitis sylvestris* around 8000 years ago (Myles et al. 2011). Due to domestication, valuable varieties of this specie have spread throughout the Mediterranean, mostly by vegetative propagation. Because of vegetative propagation and the consequent lack of evolution, *Vitis vinifera* is generally susceptible to many pests, and its cultivation requires grafting onto rootstocks resistant to Phylloxera (a pathogenic insect imported from North America). It also needs to be sprayed frequently with large amounts of pesticides, in particular against fungal pathogens not native to Europe. Pollution due to pesticides is related, inter alia, with an increased risk of developing neurodegenerative diseases such as Alzheimer's and Parkinson's disease (Zaganas et al. 2012, Hayden et al. 2010, Baldi et al. 2003). The recurrent use of pesticides is becoming environmentally and economically unsustainable

and could cause concern in densely populated areas such as western Europe, where France, Italy and Spain together produce most of the world grapes and represent the main winemakers in the world (OIV annual report, 2014). Moreover, in some cultivated areas, massive use of fungicides is leading to fungicide resistance by powdery mildew (Gadoury et al. 2012).

The main grape pathogens were imported to Europe from America: according to Levadoux (1966) Oidium arrived in 1852, Phylloxera in 1868, powdery mildew in 1876 and "black rot" in 1885. Since then, breeders (especially the French), have focused on the creation of resistant hybrids of *Vitis vinifera* and some American *Vitis* species, commonly called French-American hybrids (This et al. 2006). In the 1950s, as a result of chemical pesticides able to quickly resolve pathogenic infections, the interest of grape producers in hybrids largely evaporated, as these hybrids are generally less productive and of lower quality than pathogen-susceptible *Vitis vinifera* varieties.

The recent problems caused by pesticide pollution and pesticide-resistant pathogens have again stimulated research into hybrid varieties. Indeed, inter-specific hybrids may represent a long-term and eco-friendly solution to pathogenic infections, being in many cases the result of stringent selection processes aimed at producing varieties resistant to multiple pathogens. Starting from French breeders' material, since the 1970s new hybrids have been established, especially in Hungary and eastern Germany (e.g. PIWI grapes), a few of which have been registered for wine production. Nevertheless, during the breeding process, many of the experimental hybrids carry the defects originating from their American parents, such as poor sugar content, "foxy" taste (Sale and Wilson, 1926, Acree et al. 1990), offset flavors (Sun et al. 2011) and poor tannin content and availability (Harbertson et al. 2008). The production of quality hybrids therefore requires decades, in order to allow the breeder to further eliminate the undesired traits and obtain marketable varieties. Undesired characters can be perceived only in some vintage making it particularly difficult to spot the problem in the resulting wine.

According to Myles (2013) and Borneman et al. (2013) the recent advances in grape genomics (Jaillon et al. 2007, Velasco et al. 2007), genotypization (Myles, 2011; Emanuelli et al. 2013), proteomics, metabolomics (Mattivi et al. unpublished data), and pathogen-host interaction (Peressotti et al. 2010, Rouxel et al. 2013) allow breeders to obtain new set of varieties through hybridization or GMOs. It is now possible to take advantage of the huge number of markers published and the new technologies available. Wild grapes represent the main source of variability in the *Vitis* genus, but they have been little studied in comparison to *Vitis vinifera*. The first phenotypic studies on wild grapes date back to the 19th century and the first decades of the 20th century (Ravaz, 1902; Galet, 1952). The development of SSR markers for *Vitis* genotypization has been achieved only recently (Goto-

Yamamoto et al. 2013), while comparative targeted studies of the metabolites in *Vitis* germplasm berries have only been performed by Liang et al. (2012a, 2012b).

Conversely, explorative untargeted analysis of the metabolomes of wild grapes is lacking. From our point of view, it is necessary to extend knowledge about the metabolites present in wild grape berries and try to understand which undesirable (and desirable) traits are present in these species. It may be the case that wild grapes have metabolic pathways not present in *Vitis vinifera*, producing "interesting" classes of metabolites. For example, hydrolysable tannins are present in *Vitis rotundifolia* (Lee et al. 2005, Sandhu et al. 2010) and lignans are accumulated in *Vitis thunbergii* (Tung et al. 2011). These classes of metabolites might be present in other species, or different classes might be accumulated in unstudied *Vitis*. It would be very "interesting" to transfer such desirable traits to new hybrids.

In the "Fondazione Edmund Mach", the institute where this research was carried out, we have a huge collection of grape varieties, including inter-specific hybrids and wild grapes, recently genotypizated by Emanuelli et al. (2013) and uploaded to the *Vitis international variety catalogue* repository (www.vivc.de). In this "proof of principles" study, the aim was to study comparatively the metabolic profiles of the berry tissue of four different American wild grapes as compared to seven famous *Vitis vinifera* varieties, through untargeted LC-MS analysis, with three different goals: 1) to find differences in the metabolomes of the American *Vitis* and *Vitis vinifera*. 2) to evaluate their importance in terms of grape quality (especially for wine production). 3) to evaluate the genetic basis of the differences found. Three interspecific hybrids were included in the study. Two of them are first generation inbreeds (41B and Isabella), while the third, NERO, is a PIWI variety obtained after numerous backcrosses with *vinifera* material. The inclusion of these three varieties was targeted at understanding how the differences found between the two groups are transferred to the hybrids. The inclusion of Nero, a variety registered for vinification in 1993, made it possible to understand whether selection performed without the use of molecular markers was able to eliminate all the undesirable traits and keep the desirable ones.

7.2 Materials and methods

7.2.1 Reagents:

Acetonitrile, methanol and formic acid of LC-MS grade were purchased from SIGMA-Aldrich (Milan, Italy). Water was Milli-Q grade. Phloroglucinol was purchased from SIGMA-Aldrich. The entire standard set was purchased from SIGMA-Aldrich or Extrasynthese, and injected in the same conditions described by Theodoridis et al. (2012). Each standard spectrum was manually extracted and used to automatically assign the name to the compounds from the sample analysis. The whole standard set used in this experiment has already been reported by Shahaf et al. (2013), in their supplementary materials.

7.2.2 Sample preparation:

Grapes at technical maturity (18° brix) were collected from the germplasm collection of the "Fondazione Edmund Mach" at San Michele all'Adige (TN) Italy. The skin, pulp and seeds of fresh berries were manually separated, ground under liquid nitrogen and stored at -80° C until analysis. A list of the grapes analyzed in this work is available in Table 1.

Extraction for untargeted analysis was performed according to the method of Theodoridis et al. (2012), slightly modified. Briefly, one gram of ground tissue was extracted with methanol/water/chloroform 2/1/2 using 0.1% of formic acid, vortexed, sonicated for 10 minutes and agitated in an orbital shaker for 15 minutes and then centrifuged at 5000 rpm per 5 minutes; the upper organic phase was collected and filtered through 0.22 μm PTFE "WHATMAN" filters before injection.

Sample preparation for the quantification of the condensed tannin fraction was performed using the protocol described by Fortes Gris et al. (2011). Briefly, 1 gram of tissue was extracted and dried in a Rotavapor, and then reconstituted with 10 ml of Water. The water extract was loaded onto a C18 Sep-pak, washed with 40 ml of water and eluted with 30 ml of methanol. The elute was dried in a Rotavapor and reconstituted with 1 ml of methanol. 900 μl of this extract was diluted 1:1 with methanol, filtered and injected into the LC-MS system. The remaining 100 μl of the elute were added to 100 μl of phloroglucinol (100g/L) at a reaction temperature of 50°. After 20 minutes, the reaction

was stopped with 1 ml of sodium acetate (40 mM), diluted 1:5 with water/methanol and injected for targeted tannin analysis.

Accession name	Short name	Original pedigree	Species	Class	Country of origin
Merlot	MER	Magdeleine noire des Charentes x Cabernet Franc	*Vitis Vinifera*	Vinifera	France
Moscato Rosa	MOR	unknown	*Vitis Vinifera*	Vinifera	Greece
Gewürztraminer	GWT	unknown	*Vitis Vinifera*	Vinifera	Italy
Moscato ottonel	MOT	Chasselas x Muscat d'Eisenstadt	*Vitis Vinifera*	Vinifera	France
Iasma Eco 3*	ECO	Moscato ottonel x Malvasia Bianca di Candia	*Vitis Vinifera*	Vinifera	Italy
Riesling	RIE	(*Vitis* Sylvestris x Traminer)(?) x Heunisch weiss	*Vitis Vinifera*	Vinifera	Germany
Sauvignon Blanc	SAU	unknown	*Vitis Vinifera*	Vinifera	France
Nero	NER	Eger 2 x Gardonyi Geza	Seibel derivative hybrid	Hybrid	Hungary
Isabella	ISA	*Vitis Vinifera x Vitis Labrusca*	American/European Hybrid	Hybrid	USA
Millardet et Grasset 41 B	41B	Chasselas x *Vitis berlandieri*	American/European Hybrid	Hybrid	France
Vitis Cinerea	VCI	*Vitis Cinerea Engelmann*	*Vitis Cinerea*	American	USA
Vitis Californica	VCA	*Vitis Californica*	*Vitis Californica*	American	USA
Vitis Arizonica Texas	VAT	*Vitis Arizonica Engelmann*	*Vitis Arizonica Texas*	American	USA
Kober 5 BB	K5BB	*Vitis Berlandieri x Vitis Riparia*	*American Hybrid*	American	Germany

Table 1: Sample names: A list of the samples used in this experiment. All the samples were collected at 18° brix during the season 2013. Sample names are from the *Vitis* international variety catalogue website www.vivc.de; the institute collection is reported at the website with the number ITA362.

7.2.3 LC/MS workflow, analysis and data treatment:

Untargeted LC/MS analysis was performed using a "UPLC", interfaced to a "Synapt" (UHPLC-ESI-Q-TOF-MS), supplied by Waters, Manchester, UK, through the ESI source. We used the same LC separation method reported by Arapitsas et al. (2013): briefly, starting with 100% eluent A (0.1% formic acid in water), switching at 1.5 min to 10% of eluent B (0.1% formic acid in methanol) up to 3 min, then a gradient of up to 40% eluent B in 18 min, passing to 100% eluent B in 21 min, holding for up to 25 min and then re-equilibrating back to 100% eluent A for a total run of 28 minutes. The MS settings were the same as for Theodoridis et al. (2012).

All the samples of the three different tissues were injected together. The analysis workflow consisted of a starting queue of 1 Blank, 1 Standard mix and 3 QCs, while during the analysis we injected a standard mix every 20 analyses and a QC representative for each tissue after every 8 sample injections. The acquired spectra were directly converted into NETcdf files using Databridge software (Waters). Peak picking, alignment and principal component analysis (PCA) were performed using the internal data analysis pipeline built at our institution, using the automated data analysis pipeline recently published (Franceschi et al. 2014), based on the "XCMS" (Smith et al. 2006), "CAMERA" (Kuhl et al. 2012) and "MetaMS" (Wehrens et al. 2014) R packages.

Targeted LC-MS/MS analysis was performed using an UPLC-adapted version of the method described by Fortes Gris et al. (2011). The samples were injected into a UPLC chromatographer, interfaced to a TQ mass spectrometer (UHPLC-ESI-QqQ-MS, Waters, Manchester, UK) through an ESI source. The eluents used were: water with 0.1% formic acid (eluent A) and acetonitrile with 0.1% FA (eluent B). The gradient was as follows: starting from 95% of A, to 20% of B in 3 minutes; isocratic flow up to 4.3 minutes, and then ramping to 45% of B at 9 minutes. Then 100% of B at 11 minutes, holding up to 13 minutes, and then returning to the initial conditions of 95% of A for a total run of 17 minutes. The quantification of catechin, epicatechin, procyanidin B1, B2, gallocatechin, epigallocatechin and epicatechin gallate was done using a linear regression curve built on the injection of pure chemical standards, in the same analytical conditions through MRM. Quantification of phloroglucinol-bound flavanols was done as for epicatechin, epigallocatechin and epicatechin gallate equivalents respectively.

7.2.4 Statistical analysis:

The stability of the analysis was assured during the workflow, checking the repeatability of the QCs and STD mix injections manually by integrating representative peaks and through PCA analysis. We also checked that the CV% of the internal standards (gentisic acid, 3-indole propionic acid and 4-stilbenol) was below 25%. The metabolic differences between the different grape species were underlined using different statistical tests. Group discrimination was obtained using 1) univariate Welch's t-test analysis with corrected p-values using the "fdr" function (Vinaixa et al. 2012), setting 0.05 corrected p-value and 10 fold change as thresholds; 2) multivariate statistical analysis using SIMCA-P+12.0 software performing OPLS-da analysis, using 1 VIP-value and 0.0001 coefficients value as thresholds (Wold et al. 2001).

7.2.5 Compound identification

7.2.5.1 Database matching

The retention times of all the injected standards in the chromatographic conditions established by Theodoridis et al. (2012) were aligned with the chromatographic conditions used in this work through the website (predret.org), according to the method developed by Stanstrup and Vrhovsek (2014). The list of bio-markers was matched against the XCMS peak table using the db.comp.assign function in the R package "Chemhelper", with 30 ppm mass accuracy and a 1 min retention time window. The identified biomarkers were checked for correctness and eventually assigned as MSI level 1 (Sumner et al. 2007).

7.2.5.2 MS/MS analysis

MS/MS analysis was performed using the capabilities of the Q-TOF instrument. All the analysis was done in V mode (improved sensitivity, nominal resolution 10000); precursor ions were selected in the quadrupole and fragmented in the collision cell through a collision energy profile ranging from 20 eV to 35 eV, with leucine enkephaline as internal calibrant. Collected MS/MS spectra were queried in MetFrag (Wolf et al. 2010) and MetFusion (Gerlich and Neumann, 2013) using KEGG, Chemspider or Pubchem as compound databases, and MassBank-EU as the spectral database.

7.2.5.3 Isotopic pattern recognition and formula assignment

The isotopic patterns of all the biomarkers were obtained using the strategy described by Stanstrup et al. (2013), based on intensity weighted means of the isotopes automatically identified by CAMERA. Putative chemical formulas were obtained using these isotopic patterns through the getFormula function of the Rdisop Package in a R environment (Boecker and Liptak, 2007).

7.2.5.4 Compound Characteristics Comparison

An important identification step was "Compound Characteristics Comparison" (CCC): This method was established in our group and is currently being prepared for further publication. The idea behind the method is that the characteristics (RT, m/z, isotopic pattern, etc.) of the pseudo molecular ion of a compound detected in LC-MS are typical of its structure, so comparing the characteristics of the standards in our database with those from unknown biomarkers allows us to refer its structure to a

few possible ones (Narduzzi et al. unpublished data). A more in-depth description of the method is present in the results.

7.2.5.5 Compound identification strategy

The following identification strategy is described in Image 1. After database matching, the remaining unknown markers were subjected to MS/MS analysis. The acquired spectra were queried against MetFrag/MetFusion and the chemical structure list obtained was compared to the putative chemical formula and the CCC method and manually checked for correctness. If any of the structure matched all the characteristics, this biomarker was considered to have been identified at MSI level 2. Otherwise, if no match was found, a tentative structure was sketched, based on the putative class predicted by the CCC method, the putative formula calculated by both Rdisop and CCC method, and the losses and neutral losses observed in MS/MS analysis. The putative structure was queried again against MetFrag; when a higher score was found for the sketched structure in comparison to the previous ones, the biomarker was labelled as MSI level 2 or 3, depending on how likely the spectral matching was. The list of the identified compounds is available in supplementary table 1 (skin), 2 (seeds) and 3 (flesh).

7.2.5.6 Data mining

Hierarchical clustering analysis was conducted using the heatmap.2 function of the gplot package, scaling the data, using "Canberra" as distance function and Ward.D2 as the hierarchical clustering function, with blue as the low intensity color and red as the high intensity color. The short names indicated in the plots correspond to the identified compounds as reported in supplementary table 1(skin), 2 (seeds) and 3 (flesh).

Tissue	**Ionization mode**	**Features**	***Vv* Marker features**	***AV* Marker features**	***Vv* pseudo-molecular ions**	***AV* pseudo-molecular ions**	**Identified *Vv***	**Identified *AV***
Flesh	Positive	7726	20	117	6	24	3	21
Flesh	Negative	8906	14	88	8	18	5	14
Seeds	Positive	12607	32	44	12	10	10	9
Seeds	Negative	12624	57	52	23	5	23	2
Skin	Positive	13332	44	106	9	37	7	35
Skin	Negative	15826	36	280	13	29	12	29

Table 2: Markers and identified pseudo-molecular ions: Number of total features detected, total biomarkers in each tissue, total pseudo-molecular ions and identified compounds are reported in this table.

7.3 Results and discussion

7.3.1 Statistical analysis

Ten to fifteen thousand "features" were detected in each tissue, both in positive and negative ionization mode. Such a high number can affect statistical analysis, so we eliminated all the "features" that did not respect the 80% rule (Smilde et al. 2006), and those with an intensity threshold of below 1000 counts*sec (Vinaixa et al. 2012). In the data analysis we used both the univariate and multivariate statistical test: the tests are not mutually exclusive, while their combined use is suggested especially when the “curse of dimensionality” arises (Goodacre et al. 2007, Vinaixa et al. 2012). Ten to hundreds of "features" were recognized as markers in the different tissues. The marker lists were merged and duplicates were excluded. Features in the same CAMERA group were excluded, keeping only the supposed pseudo-molecular ions; the pseudo-molecular ions were subsequently integrated using Targetlynx® to manually supervise the data and eliminate false positives and wrong XCMS peak-picking. The results of statistical analysis are shown in Table 2.

7.3.2 Compound identification strategy

Identification of compounds is one of the main bottlenecks in untargeted LC-MS analysis, due to the multiple signals coming from the same metabolite in the MS spectrum. An automatic identification pipeline based on injected standards has been established in our institution (Franceschi et al. 2014), but the whole standard set was injected in a chromatographic gradient that we did not use in this experiment. To overcome this problem, we aligned the two chromatographic runs using standards injected in both runs and we performed retention time prediction using the website predret.org (Stanstrup and Vhrovsek, 2014). The predicted retention times were used to identify biomarker features through db.comp.assign in the R package “Chemhelper”, with 30 ppm mass accuracy and a 1 min retention time window. Further manual checking allowed us to undoubtedly assign MSI identification level 1 (Sumner et al. 2007) to all the matching biomarkers. In total, we were able to identify 32 metabolites in the flesh, 26 in the seeds and 44 in the skin at MSI level I, differently accumulated. The rest of the markers were considered “unknowns” and underwent further identification steps, as described in M&M sections 4.2, 4.3, 4.4 and 4.5.

The unknown biomarkers underwent a similar identification approach to the one published by Stanstrup et al. (2013), with the contribution of an in-house method for metabolite classification (Narduzzi et al. unpublished data). This classification method, called “compound characteristics

comparison" (CCC) is a multivariate regression model built on RT, m/z and the isotopic intensities of standard collection that predicts the main substructures present in the chemical structure of a compound. So the values of these parameters are used together to validate the prediction properties of the model and estimate the presence of particular substructures in the compound (phenolic groups, aliphatic chains, acidic groups, glycosides and so on). From the predicted substructures, the structure is further rebuilt and makes it possible to exclude all the non-matching structures proposed by MS/MS in-silico analysis.

This method gives a putative structural composition, it improves calculation of the putative chemical formula and, coupled with MS/MS data, it assigns putative structures to the unknown signals. The whole identification strategy is described in Image 1. In total we could identify 13 compounds as MSI level 2 in the flesh, 13 compounds as MSI level 2 and 6 as MSI level 3 in the seeds, 42 compounds as MSI level 2 and 1 as MSI level 3 in the skin. All the MS/MS spectra of the compounds of class 2 and 3 are showed in the supplementary table 4 at the bottom of this thesis.

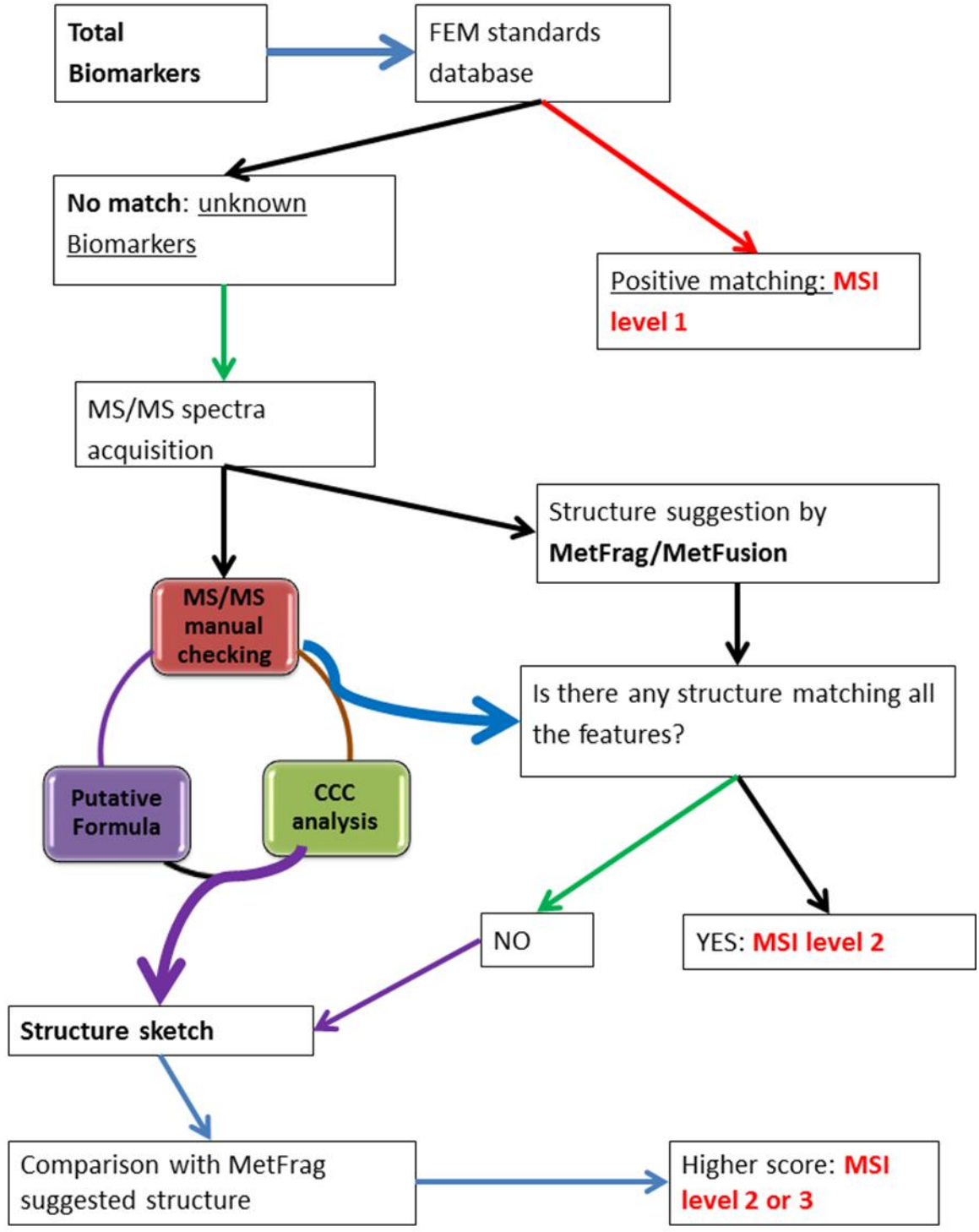


Image 1: Schematic representation of the strategy used to identify the metabolites in this paper.

7.3.3 Comparative analysis

All the identified metabolites were grouped according to their biosynthetic pathway and hierarchical clustering analysis was performed on all the samples using this data. The results are shown in Images 2 and 3 (skin and seeds respectively) and Image 4 (flesh). Procyanidins, hydrolysable tannins, aroma precursors, polar lipids, sugars, organic acids, anthocyanins, flavonols, and stilbenoids showed different accumulation in the two groups. The three latter compound classes have already been reported to be differently accumulated in *Vitis vinifera* and wild grapes in previous papers (Liang et al. 2012, Poudel et al. 2008, Hilbert et al. 2015) and will be discussed only briefly below. In the following section, we focus on the first three compound classes, as their behavior has not been reported previously and it is very interesting from the oenological and biological points of view.

7.3.3.1 Flavan-3-ols and Procyanidins

Flavan-3-ols are a very important class of polyphenols in grape and wine production, especially in their polymeric forms, as procyanidins. Procyanidins are a type of condensed tannins responsible for astringency in wine and constitute the body of the wine, together with alcohol content and some lipids. Their content is very important, especially in red wine production: lack of procyanidins means poor quality wines. Procyanidins are also recognized as the main antioxidants accumulated in grapes, and have multiple health-related characteristics (Chung et al. 1998, Bak et al. 2012). Unfortunately, the whole accumulation and polymerization pathway of procyanidins is still unknown.

Differences were found in the accumulation of Flavan-3-ols and procyanidins in the skin and the seeds of the two groups. In the skin, *Vitis vinifera* was richer in catechin, epicatechin, catechin-rhamnoside and procyanidin B3 (Image 2). It is also richer in other dimeric, trimeric and tetrameric forms of procyanidins, but this difference did not exceed the selected threshold (10-fold). Surprisingly American *Vitis* accumulated higher levels of catechin-gallate and epicatechin-gallate. *Vitis vinifera* seeds were richer in all the flavan-3-ols and procyanidins that we could identify directly in untargeted analysis (Image 3). A similar result was reported by Liang et al. (2012): *Vitis vinifera* contained a higher level of monomeric, dimeric and trimeric procyanidins in the seeds, in comparison to many different *Vitis* species except *Vitis palmata.* Fuleki et al. (1997) obtained a similar result when comparing *Vitis vinifera* with *Vitis Labrusca* seeds.

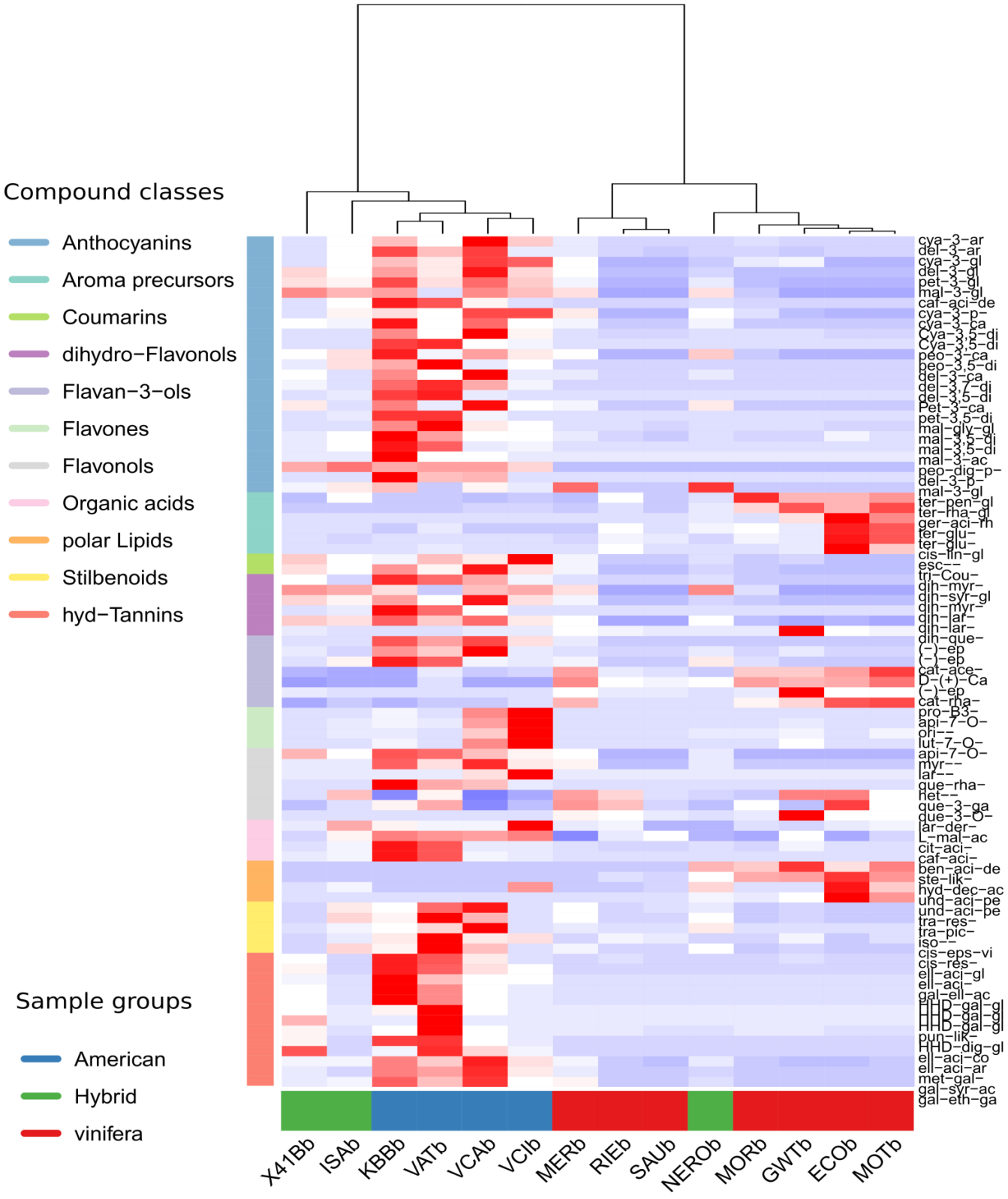


Image 2: Heat map of biomarkers from the skin. The upper legend on the left indicates the biosynthetic class of the compounds, while the bottom legend indicates the sample group.

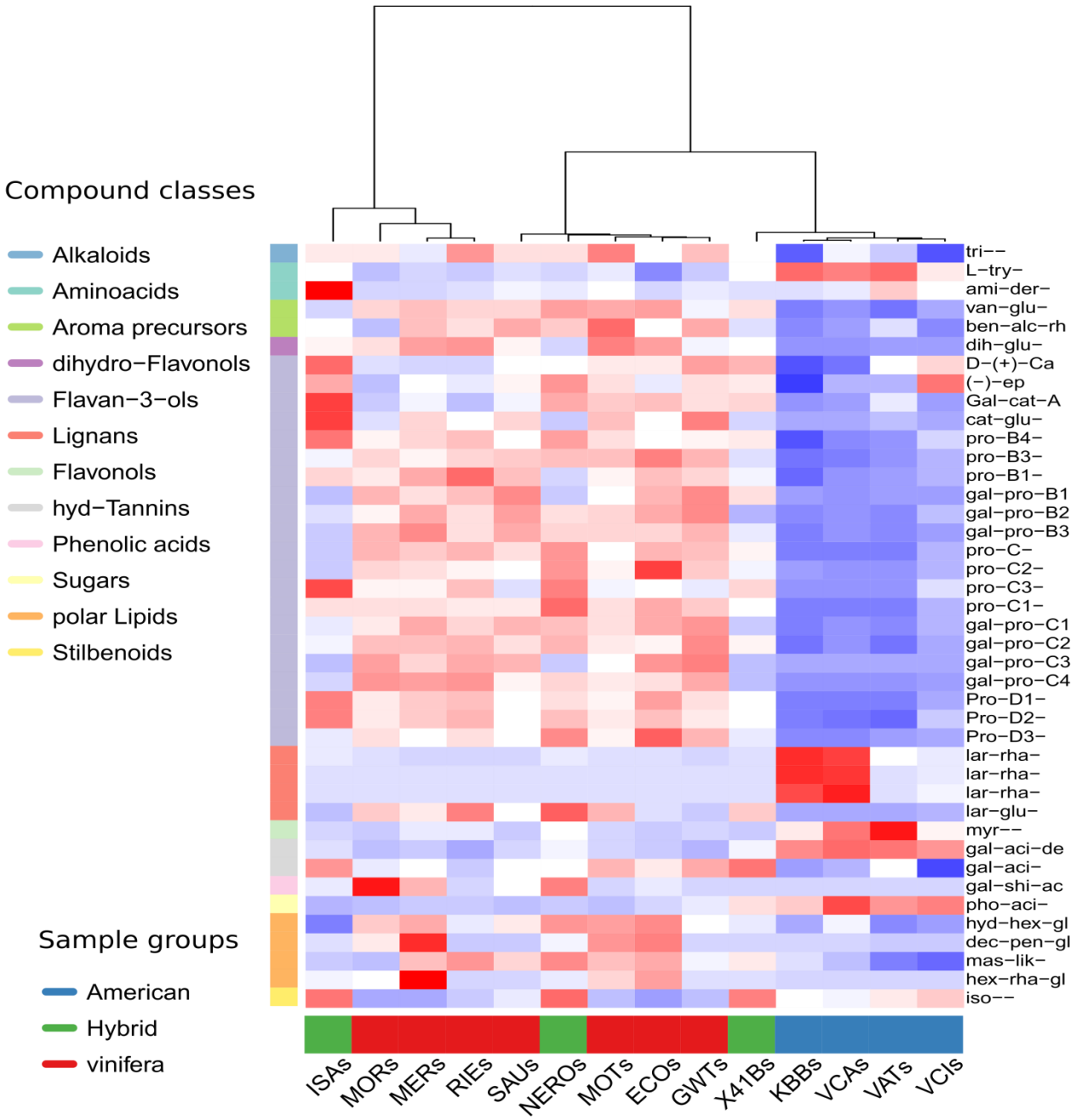


Image 3: Heat map of biomarkers from the seeds. The upper legend on the left indicates the biosynthetic class of the compounds, while the bottom legend indicates the sample group.

As stated earlier, the accumulation of procyanidins is related to many desirable characteristics; we therefore decided to perform targeted analysis of procyanidins, to quantify the difference between the two groups precisely, using the method described by Fortes Gris et al. (2011). The results obtained from targeted analysis are summarized in Image 4. *Vitis vinifera* varieties showed a higher accumulation of both free flavanols and polymeric procyanidins, especially the latter. The amount of procyanidins accumulated in *Vitis vinifera* varieties was on average 30 times higher (from 5500 to 17000 mg/Kg) in comparison to wild grapes (200-650 mg/Kg). The difference in free flavanols was about 8 times higher in *Vitis vinifera*, while American grapes accumulated a higher percentage (≈3%) of galloylated forms in comparison to *vinifera* grapes (≈1%). As expected, the mean degree of polymerization (mDP) was higher in *Vitis vinifera* (25-52 units), in comparison to wild grapes (4-17 units). The hybrid varieties showed a different pattern: Isabella and 41B had a higher percentage of galloylated units (5-7%) in comparison to all the other samples. Nero had a slightly higher level of free flavanols, polymeric procyanidins comparable to Sauvignon Blanc (5000 mg/Kg), and a mDP similar to that of the *Vitis vinifera* group. Isabella had mDP superior to that of Nero (around 39 units).

Vitis vinifera varieties accumulated more free flavanols and polymeric procyanidins in the seeds, but the difference was not so clear as in the skin. Indeed, *Vitis cinerea* accumulated a level of catechin and epicatechin comparable to the *vinifera* varieties, while it had a lower amount of dimeric forms. *Vitis cinerea's* level of polymeric procyanidins (10,000 mg/Kg) was lower than the level in *Vitis vinifera* (40,000 mg/Kg), but not as low as for the rest of the wild grapes (2000 mg/Kg). This data indicates that the accumulation of procyanidins in the seeds may vary considerably between different species. Of the hybrids, Nero showed an accumulation pattern similar to those of the *vinifera* varieties, while 41B and Isabella both had a level of polymerization degree and accumulation lying between the two groups.

Wines produced from French-American hybrids were reported to be poor in terms of mouthfeel, due to a lack of astringency, indicating the scarcity of condensed tannins (Harbertson et al. 2008). Previous authors have attributed the scarcity of condensed tannins in wines to the low availability and solubility of procyanidins from hybrid grape berries (Springer & Sacks, 2014), due to the procyanidin-binding properties of pectin and proteins. In our opinion, low extractability and low availability may together represent the main explanation for the notable difference in the detection of procyanidins in the American *Vitis* and *Vitis vinifera.* Recently, one MYB-regulatory-gene affecting procyanidin accumulation has been characterized by Huang et al. (2014). Studies of the different alleles of this gene in the *Vitis* germplasm may help to understand the procyanidin accumulation process and availability.

A more in-depth study into the reasons for such a marked difference between the two groups (especially in the skin) was outside the scope of this study, and was not investigated further. On the other hand, it is very interesting to note that the accumulation level in Nero (a recent hybrid variety) was more similar to *Vitis vinifera* than 41B and Isabella (old hybrid varieties), indicating that multiple backcrosses with *Vitis vinifera* increase the extractability and availability of procyanidins, and that their availability is a characteristic controlled by multiple factors.

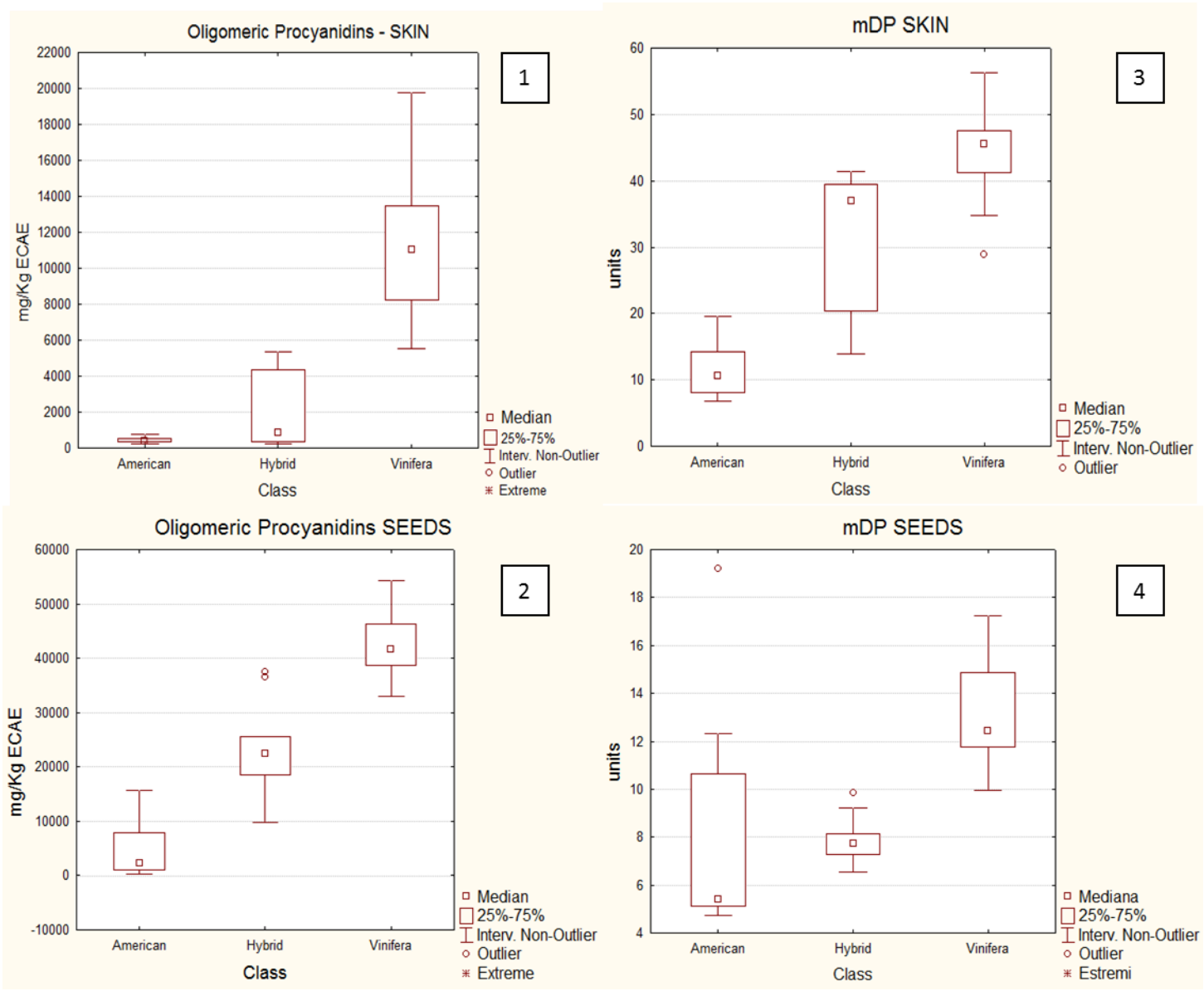


Image 4: Box plots of oligomeric procyanidin content in the skin and seeds of different classes of grapes (1 & 2) and mean degree of polymerization (mDP) of procyanidins (3 & 4). Procyanidin content is expressed as epicatechin equivalents (ECAE), while mDP is expressed as average units.

7.3.3.2 Hydrolysable tannins, precursors and their derivatives

Hydrolysable tannins are polymeric compounds constituted by gallic acid subunits bound through a glycosidic moiety. The gallic acid unit can create C-C bonds, joined to each other to create HHDP (hexahydroxydiphenoyl) subunits and/or hydrolyze their acidic moiety to form ellagic acid subunits. The different combinations of gallic acid, HHDP, ellagic acid and glycosides create the separate hydrolysable tannins. The accumulation and degradation pathway of these metabolites is still unknown.

From an oenological point of view, hydrolysable tannins are considered good quality markers. These metabolites are not found naturally in red wines from *Vitis vinifera*, indeed red wines are often aged in oak or acacia barrels, where the wine alcohol extracts lignans and hydrolysable tannins from the wood, improving wine body. Furthermore, hydrolysable tannins and their catabolic derivatives (ellagic acid and its conjugates) have many health-related properties (reviewed in Landete, 2011). For these reasons, their accumulation is welcome in wines. To our knowledge, only gallic acid, galloyl-glucose, digalloyl-glucose and ellagic acid have been reported in *Vitis vinifera* grapes. Of the *Vitis* germplasm, only *Vitis rotundifolia* has been reported to accumulate some hydrolysable tannins and derivatives in its skin and seeds (Sandhu et al. 2010, Lee et al. 2005).

In our experiment, we found that American grapes accumulated many hydrolysable tannin precursors such as gallic acid, galloyl-glucose, di-galloyl-glucose, ellagic acid, methyl-gallate, galloyl-syringic acid and galloyl-ethyl-gallate in the skin (Image 2). Oligomeric hydrolysable tannins were also found in the skin of American grapes, especially *K5bb* and *Vitis Arizonica Texas.* They accumulated in a total of eight oligomeric compounds in this class, reported here: ellagic acid-arabinoside, ellagic acid-glucoside, galloyl-ellagic acid, 3 different isomers of HHDP-galloyl-glucose, a punicalin-like compound, HHDP-digalloyl-glucose and ellagic acid conjugate 1 (MSMS spectra in Table 3). We did not observe any hydrolysable tannin, precursor or derivative signal in *Vitis vinifera* skins, except for those already reported in the literature (gallic acid, galloyl-glucose, di-galloyl-glucose and ellagic acid). Moreover, American grapes showed a concentration of ellagic acid around 1000 times higher than *vinifera* grape skin. Ellagic acid is the catabolic product of hydrolysable tannins.

We could not confirm any of the oligomeric hydrolysable tannin structures, because there are no commercial standards available on the market; the presence of many putative compounds in the same class makes us confident about the identification. Interestingly, when looking selectively for previously reported ellagitannins, we observed that *K5bb* and *Vitis Arizonica Texas* also had a m/z of 933.0692, with the same mass and many fragments in the MS/MS spectrum the same as castalagin and vescalagin, with different intensities and a very different retention time (5 and 6 min vs 12 minutes,

data not shown). In this experiment, the signal for this ion was too low and was excluded from the statistical analysis, due to the intensity threshold selected, but was clearly present in the chromatograms of *K5bb* and *Vitis Arizonica Texas*.

As shown in Image 2, of the hybrids, only *41b* showed some accumulation of these compounds in the skin, while *Nero* and *Isabella* did not accumulate it, as for *vinifera* grapes. The genetic basis of hydrolysable tannin accumulation is still unknown, so it is difficult to understand why there is a different behavior in the hybrids.

The situation became more complex when observing the accumulation of these compounds in the seeds. All the American grapes also showed accumulation of this class of compounds in the seeds but in lower intensities in comparison to the skin. Of the *vinifera* grapes, only *Moscato Rosa* showed a signal for many hydrolysable types of tannin, usually higher than that from wild grapes. This is extremely interesting, because no other vinifera showed any clear signal. These hydro-tannic units were also observed in a pilot experiment conducted last year (data not shown); in that experiment *Moscato Rosa* and the intraspecific hybrid *Pinot X Merlot* showed a clear accumulation of such compounds. Their presence in *Moscato Rosa* seeds is the reason why this class of compounds was not recognized as a marker for American grapes in seeds.

Hydrolysable tannins have been reported in numerous fruits and plants (reviewed in Arapitsas et al. 2012) but this is the first time that they have been reported in a species other than *Vitis rotundifolia* within the *Vitis* genus. From a physiological point of view, this characteristic is very interesting to us, and may be the object of a further research project.

Compound NAME	mz	RT	Spectra	intensity
HHDP-galloyl-glucose	633.073	4.5	300.999 **[**C14H6O8 -H**]**-(-11.0 ppm)	13900.00
			275.0197 **[**C13H8O7 -H**]**-(-11.6 ppm)	7814.00
			169.0142 **[**C7H5O5**]**-(-13.0 ppm)	5512.00
			302.0068 **[**C14H8O8-2H**]**-(-3.6 ppm)	2254.00
			249.0405 **[**C12H9O6**]**-(-12.0 ppm)	1721.00
			463.0518 **[**C20H17O13-2H**]**-(-19.4 ppm)	1325.00
			481.0624 **[**C20H17O14**]**-(-13.1 ppm)	1203.00
HHDP-galloyl-glucose	633.073	5.3	300.999 **[**C14H6O8 -H**]**-(-11.0 ppm)	13900.00
			275.0197 **[**C13H8O7 -H**]**-(-11.6 ppm)	6650.00
			169.0142 **[**C7H5O5**]**-(-13.0 ppm)	5154.00
			302.0068 **[**C14H8O8-2H**]**-(-3.6 ppm)	2254.00
			463.0518 **[**C20H17O13-2H**]**-(-19.4 ppm)	1971.00
			481.0624 **[**C20H17O14**]**-(-13.1 ppm)	1600.00
HHDP-galloyl-glucose	633.073	9.07	300.999 **[**C14H6O8 -H**]**-(-9.6 ppm)	7825.00
			275.0197 **[**C13H8O7 -H**]**-(-11.3 ppm)	7311.00
			463.0518 **[**C20H17O13-2H**]**-(-6.0 ppm)	2209.00
			302.0068 **[**C14H8O8-2H**]**-(-3.3 ppm)	1244.00
			481.0624 **[**C20H17O14**]**-(-1.7 ppm)	685.30
Tellimangradin I	785.089	10.47	300.999 **[**C14H6O8 -H**]**-(+3.6 ppm)	1896.00
			275.0197 **[**C13H8O7 -H**]**-(+10.4 ppm)	565.40
			463.0518 **[**C20H17O13-2H**]**-(-6.0 ppm)	284.60
Punicalin	781.0551	11.53	463.0518 **[**C20H16O13 -H**]**-(-7.3 ppm)	27040.00
			754.0659 **[**C33H22O21**]**-(9.0 ppm)	23640.00
			737.0632 **[**C33H22O20 -H**]**-(9.5 ppm)	10060.00
			299.9912 **[**C14H4O8**]**-(-34.3 ppm)	9511.00
			719.0526 **[**C33H20O19 -H**]**-(-2.5 ppm)	8817.00
			736.0553 **[**C33H20O20**]**-(5.2 ppm)	7156.00
			753.0581 **[**C33H22O21 -H**]**-(0.1 ppm)	5405.00
			735.0475 **[**C33H20O20 -H**]**-(-1.0 ppm)	3091.00
			746.0397 **[**C34H19O20 -H**]**-(6.0 ppm)	2499.00
			745.0319 **[**C34H19O20-2H**]**-(3.2 ppm)	2017.00
			763.0424 **[**C34H21O21-2H**]**-(3.8 ppm)	1964.00
Ellagic acid-arabinoside	433.0403	19.16	300.999 **[**C14H5O8**]**-(-5.6 ppm)	47720.00
			299.9912 **[**C14H5O8 -H**]**-(-5.3 ppm)	37940.00
Ellagic acid-glucoside	463.0514	15.9	299.9912 **[**C14H5O8 -H**]**-(-1.3 ppm)	121100.00
			300.999 **[**C14H5O8**]**-(2.3 ppm)	120600.00
			89.0244 **[**C3H6O3 -H**]**-(-14.6 ppm)	2690.00
			275.0197 **[**C13H8O7 -H**]**-(+2.8 ppm)	1789.00
Galloyl-ellagic acid	469.005	13.97	300.999 **[**C14H5O8**]**-(-8.3 ppm)	7047
			299.9912 **[**C14H5O8 -H**]**-(-7.3 ppm)	4859
			298.9833 **[**C14H5O8-2H**]**-(-31.4 ppm)	919.3
			271.9963 **[**C13H5O7 -H**]**-(-10.7 ppm)	802
			270.9884 **[**C13H5O7-2H**]**-(-9.6 ppm)	775
			425.015 **[**C20H9O11**]**-(5.9 ppm)	641.6
			273.0041 **[**C13H5O7**]**-(9.9 ppm)	166.4

Table 3: MS/MS spectra of hydrolysable tannins. The MS/MS spectra from *Vitis* Arizonica Texas skin are shown. Only ions identified by metFrag are reported in this table. A more complete MS/MS spectra analysis is reported in supplementary table 4.

7.3.3.3 Aroma precursors

Aroma precursors are all the volatile metabolites stored in the vacuole through glycosylation and they serve as a reservoir for biotic and abiotic stress response. Many of these glycosylated volatiles undergo hydrolysis during vinification, determining most of the wine aroma (Fernandez-Gonzales et al. 2004). For this reason, is important to study the intact aroma precursors found in the berries, to understand the aroma potential of grapes. These metabolites are mostly accumulated in the skin, and to a lesser extent in the flesh.

American grapes are often reported to have a different range of aromas, including the undesirable "foxy" taste (Acree et al. 1990) and some herbaceous flavors (Sun et al. 2011). The goal was to establish whether the putative precursors of these volatiles can be found in any of the tissues of American grapes. Surprisingly, none of the markers obtained in this work corresponded to a putative precursor of this kind of metabolite. This means that none of the MS/MS spectra could be associated with one of the known off-flavors typical of the wild American grapes.

On the other hand, it was interesting to see whether *Vitis vinifera* accumulates precursors with a pleasant aroma, such as terpenols, differentially (Image 2). Our experiment showed that *viniferas* accumulated a huge amount of glycosilated-terpenoids in the skin, while American grapes did not accumulate any. Nevertheless, high concentrations were found in Iasma Eco 3, Moscato Ottonel, Gewürztraminer and Moscato Rosa, which are commonly recognized as aromatic varieties. Aromatic is a phenotypic definition: a fruity smelling variety of grape is "aromatic". Recently, a single locus (VvDXS), encoding for a 1-deoxy-D-xylulose-5-phosphate enzyme, has been found to be responsible for the aromatic qualities of grapes (Battilana et al. 2011). This locus is involved in the accumulation of isoprenoid precursors; isoprenoids are the building blocks for mono-Terpenoids. The remaining *vinifera* varieties showed a lower accumulation of these aroma precursor volatiles, similar to American grapes. We can conclude that this difference is due more to the locus VvDXS, as reported in the literature, than to a difference between species. On the other hand, this locus was selected during domestication of the *Vitis vinifera* (Emanuelli et al. 2013), so it is common in the *vinifera* grapes and probably very rare in wild grapes. We noticed the same pattern in the flesh (Image 5), with a lower concentration in comparison to the skin, as reported in previous works (Luan & Wust, 2002). Hybrids showed a level of accumulation of these metabolites similar to that shown by non-aromatic *Vinifera* and wild grapes, indicating that they probably do not have the active allele of the VvDXS gene.

In the seeds (Image 3), we were able to identify two aroma precursors with accumulations ten times higher in *vinifera* seeds than in American seeds: vanilliyl-glucoside, and benzyl-alcohol-

rhamnosyl-glucoside. Both compounds are catabolic derivatives of the lignin production pathway. In this tissue, hybrid grapes showed similar levels to the *vinifera* grapes.

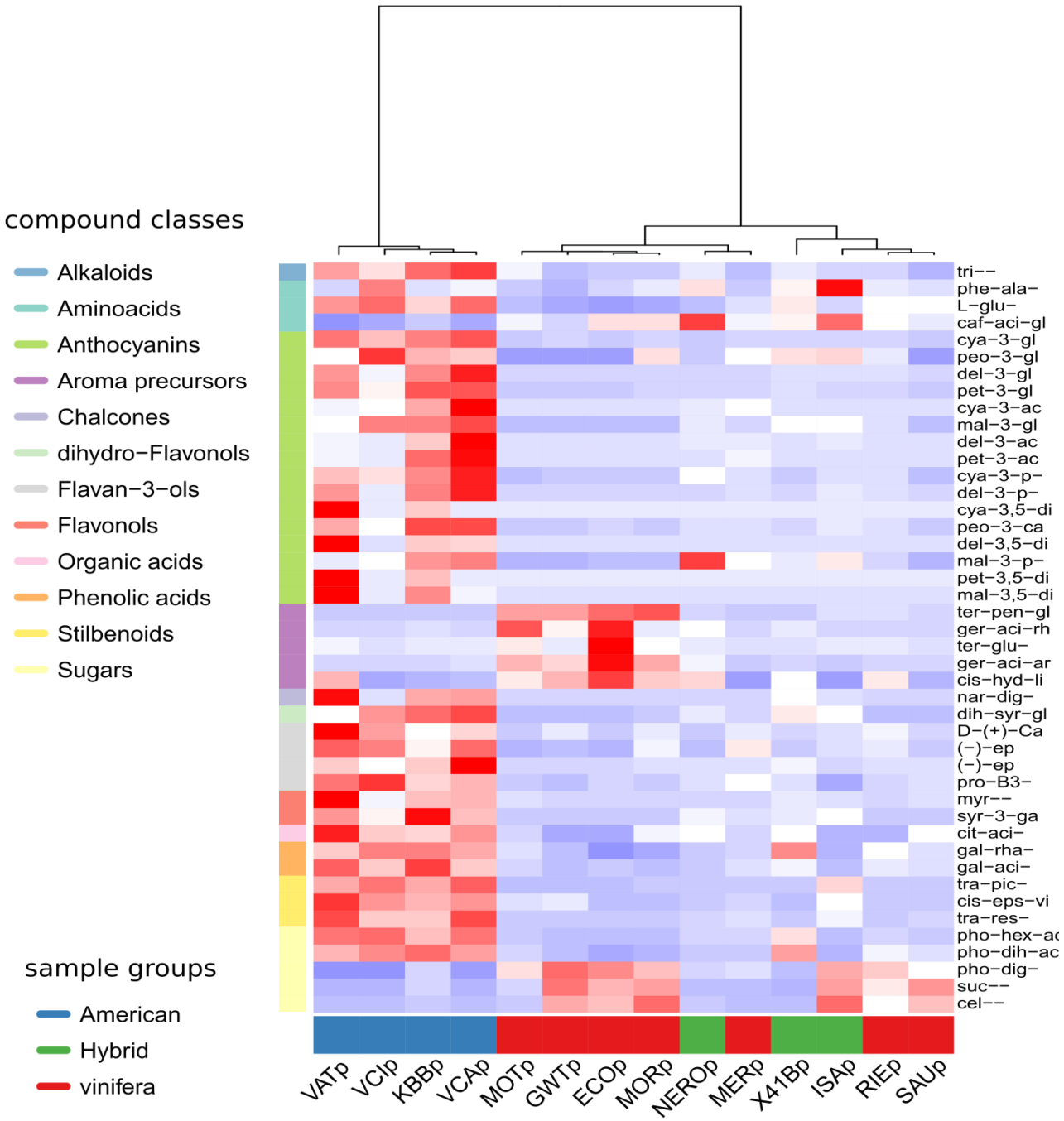


Image 5: Heatmap of biomarkers from the flesh. The upper legend on the left indicates the biosynthetic class of the compounds. While the bottom legend indicates the sample group

7.3.3.4 Anthocyanins and stilbenoids

Anthocyanins are the main pigments found in grapes, determining the visual impact of both grapes and wines. Anthocyanins are accumulated in the vacuole of the cells after glycosydation and are believed, inter alia, to defend the cells from oxidative damage due to exposure to sunlight (Matus et al. 2009). Anthocyanidins (i.e., the aglycones) cannot be found in grapes, because the aglycone form is very reactive and unstable. They share most of their metabolic pathway with flavan-3-ols, and probably compete for the substrates during the accumulation process.

A differential accumulation of anthocyanins in wild grapes and *Vitis vinifera* has been reported for many years; indeed, wild grapes tend to accumulate mostly di-glycosidic anthocyanins, while *Vitis vinifera* contains almost exclusively mono-glycosydic anthocyanins (Burns et al. 2002). Recently Yang et al. (2014) reported that the gene encoding for an 5-O-glycosyl-transferase (5-UFGT) has different alleles in *Vitis vinifera* and wild grapes, and that the alleles present in the *vinifera* grape are almost all non-functional, so they do not produce the 5-UFGT enzyme able to bind glycosides to the 5 terminal end of anthocyanins. This differential accumulation has been used historically to determine whenever a wine has been blended with non-vinifera or hybrid grapes.

In our experiment, five different basic units of anthocyaninins were found to be differently accumulated in the skin and the flesh of the two groups (Image 2 and Image 5): cyanidin, peonidin, petunidin, delphinidin and malvidin. In the skin, *vinifera* berries accumulated only anthocyanins glycosilated at position 3, sometimes with a p-coumaroyl or caffeoyl moiety attached to the glucoside; American *Vitis* was richer than *vinifera* in all the anthocyanins identified, including 3-glucosylated forms from ten to hundreds of times. Furthermore, the American *Vitis* also accumulated anthocyanins di-glycosylated in positions 3 and 5, as shown in Image 5. As expected, all the hybrids showed a lower accumulation of di-glycosilated anthocyanins than the wild American grapes.

The flesh of the *vinifera* grapes did not accumulate anthocyanins, while all the American *Vitis* showed the typical red color in their flesh, indicating the presence of a high number of anthocyanins (Image 5). Tenturier (red-fleshed) *vinifera* do exist, but they are a minority in the *vinifera* germplasm, while white-fleshed American grapes exist only in *Vitis riparia and Vitis berlandieri*. The anthocyanin accumulation pattern seems to be similar in the flesh and skin of wild grapes, but the relative amount was lower in the flesh. Of the hybrids, only NERO showed a certain degree of accumulation of anthocyanins in the flesh, due to a red layer on the external part of the flesh, but this is due to the cold storage effect more than physiological accumulation.

Stilbenoids are polyphenolic compounds that have a strong anti-fungal activity in plants and health-related characteristics in humans (e.g. resveratrol belongs to this class). Their accumulation is reported to be directly correlated with the accumulation of anthocyanins during ripening stages. As expected, due to the high concentration of anthocyanins in the skin and flesh of American grape berries, these were also richer in stilbenoids, especially *Vitis Arizonica* Texas and *Vitis Californica*, We found larger amounts of isorhapontin, cis-epsilon-viniferin, trans-resveratrol, trans-piceide, and cis-resveratrol in the skin (Image 2). Higher amounts of trans-resveratrol, trans-piceide and cis-resveratrol were also found in the flesh. As hybrid flesh is colorless (without anthocyanins), both 41b, Nero and Isabella showed very small amounts of stilbenoids in the flesh. Despite the strong anthocyanin color in their skin, the accumulation of stilbenoids in the skin of 41b and Isabella was poor, while Nero skin had a pattern similar to *Vitis* cinerea, which contains the lowest amount of the American grapes.

7.3.3.5 Flavonols and dihydro-flavonols

Flavonols are very strong antioxidants (Burda et al. 2001), and have health-related properties (reviewed in de Pasqual-Teresa et al. 2010). Flavonols are yellow pigments and determine white grape and wine color. In plants, they are believed to defend the tissues from oxidative damage by UV-B rays (Matus et al. 2009).

In our experiment, we were able to identify four basic units of flavonols differently accumulated in the two groups: quercetin, myricetin, laricitrin, and syringetin. These were present in glycosylated or rutinosylated forms. The main differences were found in the skin, where *Vinifera* tended to accumulate more quercetin-glucoside and galactoside than American grapes, while American grapes accumulated more myricetin, laricitrin and syringetin-glucoside. Nevertheless, considering grape color, and comparing the American varieties only with Merlot and Moscato Rosa, the differences were not so big, being only 2.5 times more concentrated. As reported by Mattivi et al. (2006), white grapes do not accumulate flavonols trihydroxylated in the B ring (myricetin, laricitrin and syringetin), while red ones do. So this difference seems to be due mostly to grape color rather than to grape species.

Interestingly, we could observe a similar pattern for dihydro-flavonols. This class of compounds has only recently been reported in grapes (De Rosso et al. 2014), and seemed to follow the same pattern as flavonols in our results, with the American grapes, *Merlot* and *Moscato Rosa* accumulating the trihydroxylated forms (dihydro-syringetin, dihydro-laricitrin and dihydro-myricetin) while the remaining *vinifera* accumulated only dihydro-quercetin. As dihydro-flavonols are the direct precursors of flavonols and considering that the hydroxylation of the B ring happens earlier in the metabolic

pathway (Winkel-Shirley, 2001), a similar pattern for dihydro-flavonols and flavonols would seem to be legitimate. Nevertheless, further investigations with a wider population of red and white grapes (similar to the experiment performed by Mattivi et al. 2006) are necessary to confirm these important results.

7.3.3.6 Other identified metabolites

Many other classes of metabolites were found to accumulate differently in wild American grapes and *Vitis vinifera.* Some sugar precursors, polar lipids, (putative) lignans, flavones and some organic acids showed a different accumulation in the two groups. Nevertheless, it was not possible to determine the difference found in these classes of metabolites, because many of the identifications were putative, under-represented and with a non-unique accumulation pattern.

The putative identified polar lipids were all accumulated in larger amounts in *vinifera* tissues. Hydroxy-hexanoate-glucoside, decyl-pentosyl-glucoside, mascaroside-like and hexyl-rhamnosyl-glucoside accumulated more in the seeds of the *vinifera* group than in the wild grapes, with Nero having similar levels to the *vinifera* group. A similar pattern was observed in the skin for the compounds hydroxy-decanoic acid-pentosyl-glucoside, undecanoic acid-pentosyl-glucoside and undecadioic acid-pentosyl-glucoside, with Nero again accumulating similar levels to those of *vinifera*.

Of the remaining compounds, lignans stand out as an interesting class: two signals, 505.2080 and 491.1940, were found exclusively in the American group, specifically in *Vitis Arizonica Texas* and *Vitis cinerea.* Their identification is merely putative, but their fragmentation spectra were very similar to lariciresinol with a xyloside and rhamnoside moiety respectively (data not shown). These metabolites have been previous identified in *Vitis thunbergii* (Tung et al. 2011), and are considered to be a very desirable marker, because of health benefits related to lignans.

7.4 Concluding remarks

The aim of this work was to find the metabolic differences that make *vinifera* suitable for wine production, in comparison to some American *Vitis* that are not ideal for quality wine production. The results clearly show that method adopted (untargeted analysis), allowed us to have a broad picture of many different metabolic pathways in the three different berry tissues. Having an overall view of metabolites, especially in phenolics, aroma precursors and acids, also allowed us to correlate the data to the specific metabolic pathways, in some cases revealing different metabolic pathways in the two

groups, speculating on pathway regulation and confirming already known metabolic differences between the species. The presence of hybrid grapes also improved knowledge about specific compound accumulation and behavior in different hybrids. As shown in the hierarchical clustering of images 2, 3 and 5, hybrids seem to have intermediate characteristics in relation to the two groups, with Nero (as expected) being the most similar in all tissues, and conversely 41B the most American-like, especially in terms of the skin and seeds. This work is a key step in our efforts to build up a grape berry metabolome database including metabolite localization in the different berry tissues.

References chapter 7

1. Acree, T. E.; Lavin, E. H In Flavour Science and Technology; 1990; 23, 49-52.
2. Arapitsas, P. (2012). Hydrolyzable tannin analysis in food. Food Chemistry, 135(3), 1708–17. doi:10.1016/j.foodchem.2012.05.096
3. Arapitsas, P., Speri, G., Angeli, A., Perenzoni, D., & Mattivi, F. (2014). The influence of storage on the "chemical age" of red wines. Metabolomics, 10(5), 816–832. doi:10.1007/s11306-014-0638-x
4. Bak, M. J., Jun, M., & Jeong, W. S. (2012). Procyanidins from wild grape (*Vitis* amurensis) seeds regulate ARE-mediated enzyme expression via Nrf2 coupled with p38 and PI3K/Akt pathway in HepG2 cells. International Journal of Molecular Sciences, 13, 801–818. doi:10.3390/ijms13010801
5. Baldi, I. (2003). Neurodegenerative Diseases and Exposure to Pesticides in the Elderly. American Journal of Epidemiology, 157(5), 409–414. doi:10.1093/aje/kwf216.
6. Battilana, J., Emanuelli, F., Gambino, G., Gribaudo, I., Gasperi, F., Boss, P. K., & Grando, M. S. (2011). Functional effect of grapevine 1-deoxy-D-xylulose 5-phosphate synthase substitution K284N on Muscat flavour formation. Journal of Experimental Botany, 62(15), 5497–508. doi:10.1093/jxb/err231
7. Böcker, S., & Liptak, Z. (2007). Algorithmica A Fast and Simple Algorithm for the Money Changing Problem 1, 413–432.
8. Borneman, A. R., Schmidt, S. a, & Pretorius, I. S. (2013). At the cutting-edge of grape and wine biotechnology. Trends in Genetics : TIG, 29(4), 263–71. doi:10.1016/j.tig.2012.10.014
9. Burda, S., & Oleszek, W. (2001). Antioxidant and Antiradical Activities of Flavonoids. Journal of Agricultural and Food Chemistry, 49(6), 2774–2779. doi:10.1021/jf001413m
10. Burns, J., Mullen, W., Landrault, N., Teissedre, P. L., Lean, M. E. J., & Crozier, A. (2002). Variations in the Profile and Content of Anthocyanins in Wines Made from Cabernet Sauvignon and Hybrid Grapes. Journal of Agricultural and Food Chemistry, 50, 4096–4102.
11. Chung, K. T., Wei, C. I., & Johnson, M. G. (1998). Are tannins a double-edged sword in biology and health? Trends in Food Science and Technology, 9, 168–175. doi:10.1016/S0924-2244(98)00028-4
12. De Pascual-Teresa, S., Moreno, D. a, & García-Viguera, C. (2010). Flavanols and anthocyanins in cardiovascular health: a review of current evidence. International Journal of Molecular Sciences, 11(4), 1679–703. doi:10.3390/ijms11041679
13. De Rosso, M., Tonidandel, L., Larcher, R., Nicolini, G., Dalla Vedova, a, De Marchi, F., … Flamini, R. (2014). Identification of new flavonols in hybrid grapes by combined liquid chromatography-mass spectrometry approaches. Food Chemistry, 163, 244–51. doi:10.1016/j.foodchem.2014.04.110
14. Emanuelli, F., Lorenzi, S., Grzeskowiak, L., Catalano, V., Stefanini, M., Troggio, M., … Grando, M. S. (2013). Genetic diversity and population structure assessed by SSR and SNP markers in a large germplasm collection of grape. BMC Plant Biology, 13, 39. doi:10.1186/1471-2229-13-39

15. Fernández-González, M., & Di Stefano, R. (2004). Fractionation of glycoside aroma precursors in neutral grapes. Hydrolysis and conversion by Saccharomyces cerevisiae. LWT - Food Science and Technology, 37(4), 467–473. doi:10.1016/j.lwt.2003.11.003
16. Fortes Gris, E., Mattivi, F., Ferreira, E. A., Vrhovsek, U., Pedrosa, R. C., & Bordignon-Luiz, M. T. (2011). Proanthocyanidin profile and antioxidant capacity of Brazilian *Vitis* vinifera red wines. Food Chemistry, 126(1), 213–220. doi:10.1016/j.foodchem.2010.10.102.
17. Franceschi, P., Mylonas, R., Shahaf, N., Scholz, M., Arapitsas, P., Masuero, D., … Wehrens, R. (2014). MetaDB a Data Processing Workflow in Untargeted MS-Based Metabolomics Experiments. Frontiers in Bioengineering and Biotechnology, 2(December), 72. doi:10.3389/fbioe.2014.00072.
18. Fuleki, T., & daSilva, J. M. R. (1997). Catechin and procyanidin composition of seeds from grape cultivars grown in Ontario. Journal of Agricultural and Food Chemistry, 45, 1156–1160. doi:10.1021/jf960493k
19. Gadoury, D. M., Cadle-Davidson, L., Wilcox, W. F., Dry, I. B., Seem, R. C., & Milgroom, M. G. (2012). Grapevine powdery mildew (Erysiphe necator): a fascinating system for the study of the biology, ecology and epidemiology of an obligate biotroph. Molecular Plant Pathology, 13(1), 1–16. doi:10.1111/j.1364-3703.2011.00728.x
20. Galet P. In Precis de Ampelographie pratique. Cepables and Vignobles de France. 1952. Impr. P. Dehan, Montpellier
21. Gerlich, M., & Neumann, S. (2013). MetFusion: integration of compound identification strategies. Journal of Mass Spectrometry : JMS, 48(3), 291–8. doi:10.1002/jms.3123.
22. Goodacre, R., Broadhurst, D., Smilde, A. K., Kristal, B. S., Baker, J. D., Beger, R., … Wulfert, F. (2007). Proposed minimum reporting standards for data analysis in metabolomics. Metabolomics, 3(3), 231–241. doi:10.1007/s11306-007-0081-3
23. Goto-Yamamoto, N., Azuma, A., Mitani, N., & Kobayashi, S. (2013). SSR Genotyping of Wild Grape Species and Grape Cultivars of *Vitis* vinifera and V. vinifera ^|^times; V. labrusca. Journal of the Japanese Society for Horticultural Science, 82(2), 125–130. doi:10.2503/jjshs1.82.125
24. Harbertson, J. F., Hodgins, R. E., Thurston, L. N., Schaffer, L. J., Reid, M. S., Landon, J. L., … Adams, D. O. (2008). Variability of tannin concentration in red wines. American Journal of Enology and Viticulture, 59, 210–214.
25. Hayden, K. M., Norton, M. C., Darcey, D., Ostbye, T., Zandi, P. P., Breitner, J. C. S., & Welsh-Bohmer, K. a. (2010). Occupational exposure to pesticides increases the risk of incident AD: the Cache County study. Neurology, 74(19), 1524–30. doi:10.1212/WNL.0b013e3181dd4423
26. Hilbert, G., Temsamani, H., Bordenave, L., Pedrot, E., Chaher, N., Cluzet, S., … Richard, T. (2015). Flavonol profiles in berries of wild *Vitis* accessions using liquid chromatography coupled to mass spectrometry and nuclear magnetic resonance spectrometry. Food Chemistry, 169, 49–58. doi:10.1016/j.foodchem.2014.07.079
27. Huang, Y. F., Vialet, S., Guiraud, J. L., Torregrosa, L., Bertrand, Y., Cheynier, V., … Terrier, N. (2014). A negative MYB regulator of proanthocyanidin accumulation, identified through

expression quantitative locus mapping in the grape berry. New Phytologist, 201, 795–809. doi:10.1111/nph.12557
28. Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., … Wincker, P. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature, 449(7161), 463–7. doi:10.1038/nature06148
29. Kuhl, C., Tautenhahn, R., Bo, C., Larson, T. R., & Neumann, S. (2012). CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets. Analytical Chemistry.
30. Landete, J. M. (2011). Ellagitannins, ellagic acid and their derived metabolites: A review about source, metabolism, functions and health. Food Research International, 44(5), 1150–1160. doi:10.1016/j.foodres.2011.04.027.
31. Lee, J.-H., Johnson, J. V, & Talcott, S. T. (2005). Identification of Ellagic Acid Conjugates and Other Polyphenolics in Muscadine Grapes by HPLC-ESI-MS. Journal of Agricultural and Food Chemistry, 53, 6003–6010.
32. Levadoux, L. in La vgne et sa culture. Edition no. 10. 1966. Presses universitaires de France
33. Liang, Z., Yang, Y., Cheng, L., & Zhong, G. (2012). Characterization of Polyphenolic Metabolites in the Seeds of *Vitis* Germplasm. Journal of Agricultural and Food Chemistry, (60), 1291–1299.
34. Liang, Z., Yang, Y., Cheng, L., & Zhong, G. Y. (2012). Polyphenolic composition and content in the ripe berries of wild *Vitis* species. Food Chemistry, 132(2), 730–738. doi:10.1016/j.foodchem.2011.11.009
35. Luan, F., & Wüst, M. (2002). Differential incorporation of 1-deoxy-D-xylulose into (3S)-linalool and geraniol in grape berry exocarp and mesocarp. Phytochemistry, 60, 451–459. doi:10.1016/S0031-9422(02)00147-4
36. Mattivi, F., Guzzon, R., Vrhovsek, U., Stefanini, M., & Velasco, R. (2006). Metabolite Profiling of Grape : Flavonols and Anthocyanins. Journal of Agricultural and Food Chemistry, (54), 7692–7702.
37. Matus, J. T., Loyola, R., Vega, A., Peña-Neira, A., Bordeu, E., Arce-Johnson, P., & Alcalde, J. A. (2009). Post-veraison sunlight exposure induces MYB-mediated transcriptional regulation of anthocyanin and flavonol synthesis in berry skins of *Vitis* vinifera. Journal of Experimental Botany, 60(3), 853–867. doi:10.1093/jxb/ern336
38. Myles, S., Boyko, A. R., Owens, C. L., Brown, P. J., Grassi, F., & Aradhya, M. K. (2011). Genetic structure and domestication history of the grape. Proceedings of the National Academy of Sciences of the United States of America, 108(9), 3530–3535. doi:10.1073/pnas.1009363108/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1009363108
39. Myles, S. (2013). Improving fruit and wine: what does genomics have to offer? Trends in Genetics : TIG, 29(4), 190–6. doi:10.1016/j.tig.2013.01.006
40. Peressotti, E., Wiedemann-Merdinoglu, S., Delmotte, F., Bellin, D., Di Gaspero, G., Testolin, R., … Mestre, P. (2010). Breakdown of resistance to grapevine downy mildew upon limited deployment of a resistant variety. BMC Plant Biology, 10, 147. doi:10.1186/1471-2229-10-147

41. Poudel, P. R., Tamura, H., Kataoka, I., & Mochioka, R. (2008). Phenolic compounds and antioxidant activities of skins and seeds of five wild grapes and two hybrids native to Japan. Journal of Food Composition and Analysis, 21, 622–625. doi:10.1016/j.jfca.2008.07.003
42. Ravaz L. in Les vignes Americaines: Porte-Greffes et Producteurs-directs. Caracteres – Aptitudes. 1902. Coulet et fils Editeurs, Montpellier
43. Rouxel, M., Mestre, P., Comont, G., Lehman, B. L., Schilder, A., & Delmotte, F. (2013). Phylogenetic and experimental evidence for host-specialized cryptic species in a biotrophic oomycete. New Phytologist, 197, 251–263. doi:10.1111/nph.12016
44. Sale J. W., Wilson J. B., Jour. Agr. Res. 1926, 33, 301-10.
45. Sandhu, A. K., & Gu, L. (2010). Antioxidant capacity, phenolic content, and profiling of phenolic compounds in the seeds, skin, and pulp of *Vitis* rotundifolia (Muscadine Grapes) As determined by HPLC-DAD-ESI-MS(n). Journal of Agricultural and Food Chemistry, 58(8), 4681–92. doi:10.1021/jf904211q
46. Shahaf, N., Franceschi, P., Arapitsas, P., Rogachev, I., Vrhovsek, U., & Wehrens, R. (2013). Constructing a mass measurement error surface to improve automatic annotations in liquid chromatography/mass spectrometry based metabolomics. Rapid Communications in Mass Spectrometry, 27(21), 2425–2431. doi:10.1002/rcm.6705
47. Smilde, A. K., van der Werf, M. J., Bijlsma, S., van der Werff-van der Vat, B. J. C., & Jellema, R. H. (2005). Fusion of mass spectrometry-based metabolomics data. Analytical Chemistry, 77(20), 6729–36. doi:10.1021/ac051080y
48. Smith, C. a, Want, E. J., O'Maille, G., Abagyan, R., & Siuzdak, G. (2006). XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Analytical Chemistry, 78(3), 779–87. doi:10.1021/ac051437y
49. Springer, L. F., & Sacks, G. L. (2014). Protein-Precipitable Tannin in Wines from *Vitis* vinifera and Interspeci fi c Hybrid Grapes (*Vitis* ssp.): Di ff erences in Concentration, Extractability, and Cell Wall Binding. Journal of Agricultural and Food Chemistry, 62, 7515–7523.
50. Stanstrup J, Vrhovšek U. Comprehensive sharing and mapping of RPLC retention time information **[**Conference poster**]**. AISBM 2014 – Challenges in annotation and de novo identification of small molecules. Gif-sur-Yvette, France.
51. Stanstrup, J., Gerlich, M., Dragsted, L. O., & Neumann, S. (2013). Metabolite profiling and beyond: approaches for the rapid processing and annotation of human blood serum mass spectrometry data. Analytical and Bioanalytical Chemistry, 405(15), 5037–48. doi:10.1007/s00216-013-6954-6
52. Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. a., … Viant, M. R. (2007). Proposed minimum reporting standards for chemical analysis. Metabolomics, 3(3), 211–221.
53. Sun, Q., Gates, M. J., Lavin, E. H., Acree, T. E., & Sacks, G. L. (2011). Comparison of odor-active compounds in grapes and wines from *Vitis* vinifera and non-foxy American grape species. Journal of Agricultural and Food Chemistry, 59(Figure 1), 10657–10664. doi:10.1021/jf2026204

54. Theodoridis, G., Gika, H., Franceschi, P., Caputi, L., Arapitsas, P., Scholz, M., … Mattivi, F. (2011). LC-MS based global metabolite profiling of grapes: solvent extraction protocol optimisation. Metabolomics, 8(2), 175–185. doi:10.1007/s11306-011-0298-z
55. This, P., Lacombe, T., & Thomas, M. R. (2006). Historical origins and genetic diversity of wine grapes. Trends in Genetics : TIG, 22(9), 511–9. doi:10.1016/j.tig.2006.07.008
56. Tung, Y. T., Cheng, K. C., Ho, S. T., Chen, Y. L., Wu, T. L., Hung, K. C., & Wu, J. H. (2011). Comparison and Characterization of the Antioxidant Potential of 3 Wild Grapes-*Vitis* thunbergii,V. flexuosa, andV. kelungeusis. Journal of Food Science, 76(5), 701–706. doi:10.1111/j.1750-3841.2011.02178.x
57. Tung, Y. T., Cheng, K. C., Ho, S. T., Chen, Y. L., Wu, T. L., Hung, K. C., & Wu, J. H. (2011). Comparison and Characterization of the Antioxidant Potential of 3 Wild Grapes-*Vitis* thunbergii,V. flexuosa, andV. kelungeusis. Journal of Food Science, 76(5), 701–706. doi:10.1111/j.1750-3841.2011.02178.x
58. Vinaixa, M., Samino, S., Saez, I., Duran, J., Guinovart, J. J., & Yanes, O. (2012). A Guideline to Univariate Statistical Analysis for LC/MS-Based Untargeted Metabolomics-Derived Data. Metabolites, 2(4), 775–795. doi:10.3390/metabo2040775Wehrens, R.; Weingart, G.; Mattivi, F. J. Chromatogr. B. Analyt. Technol. Biomed. Life Sci. 2014, 966, 109–116.
59. Winkel-shirley, B. (2001). Flavonoid Biosynthesis . A Colorful Model for Genetics , Biochemistry , Cell Biology , and Biotechnology 1. Plant Physiology, 126(January 2015), 485–493.
60. Wold, S., Sjostrom, M., & Eriksson, L. (2001). PLS-regression : a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems, 58, 109–130.
61. Wolf, S., Schmidt, S., Müller-Hannemann, M., & Neumann, S. (2010). In silico fragmentation for computer assisted identification of metabolite mass spectra. BMC Bioinformatics, 11, 148.
62. Yang, Y., Labate, J. a, Liang, Z., Cousins, P., Prins, B., Preece, J. E., … Zhong, G.-Y. (2014). Multiple loss-of-function 5-O-glucosyltransferase alleles revealed in *Vitis* vinifera, but not in other *Vitis* species. TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik, 127(11), 2433–51. doi:10.1007/s00122-014-2388-6
63. Zaganas, I., Kapetanaki, S., Mastorodemos, V., Kanavouras, K., Colosio, C., Wilks, M. F., & Tsatsakis, A. M. (2013). Linking pesticide exposure and dementia: what is the evidence? Toxicology, 307, 3–11. doi:10.1016/j.tox.2013.02.002

8. Conclusions and perspectives

The hypotheses generated during the observation of the data from the grape metabolome project have been tested and verified in my thesis. I succeeded in confirming both of the experimental hypotheses regarding the identification of the volatile precursors through LC-MS analysis (chapter 5) and the individuation of further markers distinguishing the wild American grapes from the some domesticated *Vitis vinifera* varieties (chapter 7).

The strength of the method developed in chapter 5 is that it can be applied to any LC-MS method, allowing every laboratory to analyze the volatile precursors with their own chromatographic method, without the use of dedicated methodologies, which is cost-ineffective and duplicates both data acquiring and data analysis. The results of the chapter 7 are, from biological and oenological points of view, the most interesting of my thesis, demonstrating that the difference across the *Vitis* germplasm is very wide and it has not been completely exploited yet. This finding opens new possibilities in the development of grape hybrids, which may lead to novel products on the market in the next years.

On the contrary, the method developed in chapter 6 is unfortunately incomplete. The method is itself very innovative, but its integration with the state of the art methodologies for data analysis is not straightforward. It relies on features grouping, which is now the bottleneck in data analysis. It needs further development based on public databases that, at the moment, are not available to be directly used. Furthermore, the kind of relationship found is non-linear, and the development of non-linear PLSr is not an easy task.

The compounds identified in chapter 5 and chapter 7 point to a future where the entire metabolic space of the grape berry and its distribution across its tissues will be known. The application of modern spectrometric methodologies is closing the gap, allowing to dream that one day the path good-grape to good-wine will be completely understood.

Lastly, the method developed in chapter 6 pursued the classification and identification of the metabolites, reducing data analysis time and simplifying the identification process. Even though it did not reach a complete goal, it demonstrated that not all the combination of the selected parameter (X matrix) are possible, clearly indicating that different substructures of the metabolites have a recognizable influence on the molecular structure. This effect might be the case of further studies.

Mode	Chemical class	Compound name	Short-name	RT	exp MM	Cal MM	ppm	Formula	MS
-ve	organic acid	L-malic acid	L-mal-ac	1.5	134.0193	134.0215	-16.59	C4H6O5	1
-ve	tannins	methyl gallate	met-gal-	9.9	184.0370	184.0372	-0.85	C8H8O5	1
-ve	organic acid	citric acid	cit-aci-	3.2	192.0259	192.0270	-5.74	C6H8O7	1
-ve	Organic acid	Oxalyl-benzoic acid	esc--	11.6	194.0213	194.0215	-1.15	C9H6O5	2
-ve	stilbenoid	trans-resveratrol	tra-res-	19.7	228.0784	228.0786	-1.07	C14H12O3	1
-ve	tannins	ellagic acid	ell-aci-	19.9	302.0057	302.0063	-1.88	C14H6O8	1
-ve	flavan-3-ols	(-)-epigallocatechin	(-)-ep	12.0	306.0742	306.0740	0.82	C15H14O7	1
-ve	flavonol	myricetin	myr--	20.2	318.0367	318.0376	-2.77	C15H10O8	1
-ve	dihydroflavonols	dihydro-myricetin	dih-myr-	6.8	320.0526	320.0532	-1.93	C15H12O8	2
-ve	flavonol	laricitrin	lar--	20.9	332.0534	332.0532	0.62	C16H12O8	1
-ve	tannins	galloyl-syringic acid	gal-syr-ac	8.6	350.0633	350.0629	1.14	C16H14O9	2
-ve	tannins	galloyl-ethyl-gallate	gal-eth-ga	14.0	350.0633	350.0638	-1.43	C16H14O9	2
-ve	stilbenoid	trans-piceide	tra-pic-	15.8	390.1311	390.1315	-1.00	C20H22O8	1
-ve	stilbenoid	isorhapontin	iso--	16.8	420.1420	420.1420	-0.05	C21H24O9	1
-ve	flavone	apigenin-7-O-glucoside	api-7-O-	18.8	432.1050	432.1056	-1.44	C21H20O10	1
-ve	flavonol	quercetin-rhamnoside	que-rha-	16.9	448.0985	448.1006	-4.60	C21H20O13	2
-ve	flavone	orientin	ori--	16.3	448.0985	448.1006	-4.60	C21H20O11	1
-ve	flavone	luteolin 7-O glucoside	lut-7-O-	19.0	448.0998	448.1006	-1.70	C21H20O11	1
-ve	stilbenoid	cis epsilon viniferin	cis-eps-vi	20.6	454.1419	454.1416	0.58	C28H22O6	1
-ve	flavan-3-ols	(-)-epigallo-catechin gallate	(-)-ep	11.8	458.0842	458.0849	-1.48	C22H18O11	1
-ve	tannins	ellagic acid-glucoside	ell-aci-gl	15.9	464.0586	464.0591	-1.06	C20H16O13	2
-ve	tannins	galloyl-ellagic acid	gal-ell-ac	14.0	470.0122	470.0121	0.13	C21H10O13	2
-ve	dihydroflavonols	dihydro-syringetin-glucoside	dih-syr-gl	16.4	510.1354	510.1315	7.71	C30H22O8	2
-ve	tannins	HHDP-galloyl-glucose	HHD-gal-gl	4.5	634.0802	634.0806	-0.65	C27H22O18	2
-ve	tannins	HHDP-galloyl-glucose	HHD-gal-gl	5.3	634.0802	634.0806	-0.65	C27H22O18	2
-ve	tannins	HHDP-galloyl-glucose	HHD-gal-gl	9.1	634.0802	634.0806	-0.65	C27H22O18	2
-ve	tannins	punicalin-like	pun-lik-	11.5	782.0623	782.0603	2.59	C34H22O22	2
-ve	tannins	HHDP-digalloyl-glucose	HHD-dig-gl	10.5	786.0962	786.0916	5.89	C34H26O22	2
-ve	flavonol	heteronoside	het--	13.3	790.1975	790.1956	2.35	C36H38O20	2
+ve	organic acid	caffeic acid	caf-aci-	11.9	180.0437	180.0423	8.19	C9H8O4	1
+ve	coumarins	Hydroxy-benzodioxine-carboxylic acid	tri-cou-	9.8	194.0235	194.0215	9.97	C9H6O5	2
+ve	organic acid	benzoic acid derivative	ben-aci-de	11.9	222.0542	222.0528	5.99	C11H10O5	2
+ve	stilbenoid	cis-resveratrol	cis-res-	15.8	228.0796	228.0786	4.24	C14H12O4	1
+ve	tannins	ellagic acid conjugate	ell-aci-co	3.6	302.0088	302.0063	8.43	C14H6O8	2
+ve	dihydroflavonols	dihydro-myricetin	dih-myr-	10.0	320.0542	320.0532	2.96	C15H12O8	2
+ve	dihydroflavonols	dihydro-laricitrin	dih-lar-	9.2	334.0695	334.0689	1.89	C16H14O8	2
+ve	dihydroflavonols	dihydro-laricitrin	dih-lar-	13.2	334.0710	334.0689	6.49	C16H14O8	2
+ve	dihydroflavonol	dihydrosyringetin	cat-ace-	11.8	348.0860	348.0845	4.33	C17H16O8	2
+ve	anthocyanin	cyanidin-3-arabinoside	cya-3-ar	15.6	419.0973	419.0978	-1.26	C20H19O10	2
+ve	flavone	apigenin-7-O-glucoside	api-7-O-	18.7	432.1067	432.1056	2.50	C21H20O10	1
+ve	tannins	ellagic acid-arabinoside	ell-aci-ar	19.1	434.0499	434.0485	3.24	C19H14O12	2
+ve	anthocyanin	delphinidin-3-arabinoside	del-3-ar	13.9	435.0944	435.0927	3.72	C20H19O11	2
+ve	anthocyanin	cyanidin 3-glucoside	cya-3-gl	12.6	449.1087	449.1078	2.00	C21H21O11	1

+ve	anthocyanin	delphinidin 3-glucoside	del-3-gl	11.1	465.1036	465.1033	0.62	C21H21O12	1
+ve	anthocyanin	petunidin-3-glucoside	pet-3-gl	13.2	479.1192	479.1190	0.52	C22H23O12	1
+ve	anthocyanin	malvidin-3-glucoside	mal-3-gl	14.7	493.1349	493.1346	0.61	C23H25O12	1
+ve	anthocyanin	caffeic acid derivative	caf-aci-de	11.9	511.1463	511.1470	-1.32	C9H8O4	3
+ve	anthocyanin	cyanidin 3-p-coumaroyl-glucoside	cya-3-p-	20.2	595.1462	595.1452	1.74	C30H27O13	1
+ve	anthocyanin	cyanidin 3 caffeoyl-glucoside	cya-3-ca	20.0	611.1425	611.1401	3.92	C30H27O14	1
+ve	anthocyanin	Cyanindin 3,5 diglucoside	Cya-3,5-di	6.5	611.1618	611.1612	0.99	C27H31O16	1
+ve	anthocyanin	Cyanindin 3,5 diglucoside	Cya-3,5-di	8.9	611.1627	611.1612	2.44	C27H31O16	1
+ve	anthocyanin	peonidin 3-caffeoyl-glucoside	peo-3-ca	20.2	625.1572	625.1557	2.41	C31H29O14	1
+ve	anthocyanin	peonidin 3,5 diglucoside	peo-3,5-di	10.7	625.1777	625.1769	1.34	C28H33O16	1
+ve	anthocyanin	delphinidin 3-caffeoyl-glucoside	del-3-ca	18.9	627.1370	627.1350	3.20	C30H27O15	1
+ve	anthocyanin	delphidin 3,7 diglucoside	del-3,7-di	8.4	627.1573	627.1561	1.86	C27H31O17	1
+ve	anthocyanin	delphidin 3,5 diglucoside	del-3,5-di	7.6	627.1582	627.1561	3.31	C27H31O17	1
+ve	anthocyanin	Petunidin 3-caffeoyl-glucoside	Pet-3-ca	19.8	641.1542	641.1506	5.55	C31H29O15	1
+ve	anthocyanin	petunidin 3,5 diglucoside	pet-3,5-di	9.4	641.1734	641.1718	2.51	C28H33O17	1
+ve	anthocyanin	malvidin-glycosyl-glucoside	mal-gly-gl	8.8	651.1562	651.1561	0.06	C29H31O17	2
+ve	anthocyanin	malvidin-3,5-diglucoside	mal-3,5-di	11.6	655.1880	655.1874	0.88	C29H35O17	1
+ve	anthocyanin	malvidin-3,5-diglucoside	mal-3,5-di	11.5	655.1888	655.1874	2.10	C29H35O17	1
+ve	anthocyanin	malvidin 3-acetoyl-5 diglucoside	mal-3-ac	14.0	697.2012	697.1980	4.59	C31H37O18	1
+ve	anthocyanin	peonidin-diglucoside-p-coumaroyl	peo-dig-p-	19.7	771.2170	771.2136	4.33	C37H39O18	2
+ve	anthocyanin	delphinidin 3 p-coumaroyl- 5 diglucoside	del-3-p-	17.4	773.1954	773.1980	-3.43	C36H41O19	1
-ve	flavan-3-ols	D-(+)-Catechin	D-(+)-Ca	9.6	290.0782	290.0790	-2.89	C15H14O6	1
-ve	flavan-3-ols	(-)-epicatechin	(-)-ep	12.6	290.0784	290.0790	-2.20	C15H14O6	1
-ve	dihydroflavonol s	dihydro-quercetin	dih-que-	18.3	304.0579	304.0583	-1.35	C15H12O7	1
-ve	polar lipid	steroid like	ste-lik-	21.0	354.2408	354.2406	0.37	C20H34O5	3
-ve	flavan-3-ols	catechin-rhamnoside	cat-rha-	12.2	436.1361	436.1369	-1.92	C21H24O10	2
-ve	aroma precursor	terpenyl-pentosyl-glucoside	ter-pen-gl	21.2	448.2315	448.2308	1.68	C21H36O10	2
-ve	aroma precursor	terpenyl-rhamnosyl-glucoside	ter-rha-gl	21.0	462.2089	462.2101	-2.69	C21H34O11	2
-ve	flavonol	quercetin-3-galactoside	que-3-ga	19.4	464.0950	464.0955	-1.03	C21H20O12	1
-ve	aroma precursor	geranic acid-rhamnosyl-glucoside	ger-aci-rh	21.1	476.2272	476.2258	3.02	C22H36O11	1
-ve	flavonol	quercetin 3-O-glucuronide	que-3-O-	19.0	478.0739	478.0747	-1.76	C21H18O13	1
-ve	flavonol	laricitrin-derivative	lar-der-	18.3	518.1033	518.1060	-5.29	C24H22O13	2
-ve	flavan-3-ols	procyanidin B3	pro-B3-	8.3	578.1419	578.1424	-0.91	C30H26O12	1
+ve	aroma precursor	terpendiol-glucoside	ter-glu-	17.6	332.1842	332.1835	2.04	C16H28O7	3
+ve	aroma precursor	terpendiol-glucoside	ter-glu-	20.3	332.1847	332.1835	3.46	C16H28O7	2
+ve	polar lipid	hydroxy decanoic acid pentosyl-glucoside	hyd-dec-ac	12.3	482.2375	482.2363	2.44	C21H38O12	2
+ve	aroma precursor	cis-linalyloxide-glycosyl-glucoside	cis-lin-gl	14.7	494.2374	494.2363	2.16	C22H38O12	1
+ve	anthocyanin	malvidin-3-glucoside methyl acetate	mal-3-gl	19.6	535.1442	535.1452	-1.87	C25H27O13	1

Supplementary table 1: a list of the identified compounds from the grape skin in the chapter 7.

Mode	Chemical class	Compound name	Short-name	RT	exp MM	Cal MM	ppm	Formula	MSI
-ve	aminoacid	L-tryptophan	L-try-	8.3	204.0892	204.0899	-3.43	C11H12N2O2	1
-ve	sugar	phosphohexonoic acid	pho-aci-	1.5	276.0272	276.0246	9.42	C6H13O10P	2
-ve	flavan-3-ols	D-(+)-Catechin	D-(+)-Ca	9.5	290.0782	290.0790	-2.76	C15H14O6	1
-ve	flavan-3-ols	(-)-epicatechin	(-)-ep	12.6	290.0782	290.0790	-2.76	C15H14O6	1
-ve	flavonol	myricetin	myr--	20.2	318.0371	318.0376	-1.42	C15H10O8	1
-ve	tannins	gallic acid derivative	gal-aci-de	15.0	464.0720	464.0743	-4.96	C24H16O10	3
-ve	flavonoid	lariciresinol-xyloside	fla-glu-	18.9	492.2037	492.1995	8.56	C25H32O10	3
-ve	flavonoid	lariciresinol-rhamnoside	fla-glu-	17.8	506.2132	506.2152	-3.95	C26H34O10	3
-ve	flavonoid	Lariciresinol-rhamnoside	fla-glu-	18.3	506.2153	506.2152	0.20	C26H34O10	3
+ve	aminoacid	aminoacid derivative	ami-der-	16.7	373.1534	373.1525	2.41	C20H23NO6	3
+ve	stilbenoid	isorhapontin	iso--	18.2	420.1403	420.1420	-4.05	C21H24O9	1
-ve	phenolic acid	gallic acid	gal-aci-	6.8	170.0202	170.0215	-7.65	C7H6O5	1
-ve	polar lipid	hydroxy-hexanoate-glucoside	hyd-hex-gl	11.4	294.1309	294.1315	-2.04	C12H22O8	2
-ve	phenolic acid	galloyl-shikimic acid	gal-shi-ac	7.1	326.0631	326.0638	-2.02	C14H14O9	2
-ve	flavan-3-ols	Galloyl-catechin A	Gal-cat-A	20.5	440.0733	440.0743	-2.27	C22H16O10	3
-ve	flavan-3-ols	catechin-glucoside	cat-glu-	10.1	452.1301	452.1319	-3.98	C21H24O11	2
-ve	polar lipid	decyl-pentosyl-glucoside	dec-pen-gl	20.8	466.2774	466.2778	-0.86	C22H42O10	2
-ve	polar lipid	mascaroside-like	mas-lik-	18.9	524.2249	524.2258	-1.72	C26H36O11	3
-ve	flavan-3-ols	procyanidin B4	pro-B4-	10.5	578.1423	578.1424	-0.17	C30H26O12	1
-ve	flavan-3-ols	procyanidin B3	pro-B3-	8.2	578.1426	578.1424	0.35	C30H26O12	1
-ve	flavan-3-ols	procyanidin B1	pro-B1-	9.3	578.1427	578.1424	0.52	C30H26O12	1
-ve	flavan-3-ols	galloyl-procyanidin B	gal-pro-B1	11.0	730.1539	730.1534	0.68	C37H30O16	1
-ve	flavan-3-ols	galloyl-procyanidin B	gal-pro-B2	11.8	730.1544	730.1534	1.37	C37H30O16	1
-ve	flavan-3-ols	procyanidin C	pro-C-	5.3	866.2075	866.2058	1.96	C45H38O18	1
-ve	flavan-3-ols	procyanidin C2	pro-C2-	9.2	866.2075	866.2058	1.96	C45H38O18	1
-ve	flavan-3-ols	procyanidin C3	pro-C3-	12.0	866.2078	866.2058	2.31	C45H38O18	1
-ve	flavan-3-ols	procyanidin C1	pro-C1-	8.5	866.2084	866.2058	3.00	C45H38O18	1
-ve	flavan-3-ols	galloyl-procyanidin C	gal-pro-C1	14.0	1018.2202	1018.2168	3.34	C52H42O22	1
-ve	flavan-3-ols	galloyl-procyanidin C	gal-pro-C2	11.5	1018.2212	1018.2168	4.32	C52H42O22	1
-ve	flavan-3-ols	galloyl-procyanidin C	gal-pro-C3	16.5	1018.2212	1018.2168	4.32	C52H42O22	1
-ve	flavan-3-ols	galloyl-procyanidin C	gal-pro-C4	17.2	1018.2222	1018.2168	5.30	C52H42O22	1
-ve	flavan-3-ols	Procyanidin D1	Pro-D1-	10.1	1154.2762	1154.2692	6.06	C60H50O24	1
-ve	flavan-3-ols	Procyanidin D2	Pro-D2-	12.2	1154.2762	1154.2692	6.06	C60H50O24	1
-ve	flavonoid	Procyanidin D	Pro-D3-	8.1	1156.2912	1156.2848	5.53	C60H52O24	1
+ve	alkaloid	trigonelline	tri--	1.5	137.0475	137.0477	-1.46	C7H7NO2	1
+ve	aroma precursor	vanilliyl-glucoside	van-glu-	11.4	316.1127	316.1158	-9.81	C14H20O8	2
+ve	polar lipid	hexyl-rhamnosyl-glucoside	hex-rha-gl	20.0	394.2203	394.2203	-0.35	C18H34O9	2
+ve	aroma precursor	benzyl-alcohol-rhamnosyl-glucoside	ben-alc-rh	12.0	416.1688	416.1682	1.44	C19H28O10	1
+ve	dihydro-flavonol	dihydrokaempferol-glucoside	dih-glu-	12.4	450.1176	450.1162	3.11	C21H22O11	2
+ve	flavan-3-ols	procyanidin B2	pro-B2-	12.2	578.1433	578.1424	1.56	C30H26O12	1
+ve	flavan-3-ols	procyanidin B3	pro-B3-	9.2	578.1436	578.1424	2.08	C30H26O12	1
+ve	flavan-3-ols	procyanidin B4	pro-B4-	10.4	578.1442	578.1424	3.11	C30H26O12	1
+ve	flavan-3-ols	galloyl-procyanidin B	gal-pro-B3	17.7	730.1569	730.1534	4.79	C37H30O18	1

Supplementary table 2: a list of the identified compounds from the grape seeds in the chapter 7.

Mode	Chemical class	Compound name	Short-name	RT	exp MM	Cal MM	ppm	Formula	MSI
-ve	organic acid	citric acid	cit-aci-	1.6	192.0252	192.0270	-9.37	C6H8O7	1
-ve	sugar	phosphohexonoic acid	pho-hex-ac	1.5	276.0280	276.0246	12.32	C6H13O10P	2
-ve	flavan-3-ols	D-(+)-Catechin	D-(+)-Ca	9.6	290.0784	290.0790	-2.07	C15H14O6	1
-ve	flavan-3-ols	(-)-epicatechin	(-)-ep	12.6	290.0785	290.0790	-1.72	C15H14O6	1
-ve	flavan-3-ols	(-)-epigallocatechin	(-)-ep	9.4	306.0735	306.0740	-1.74	C15H14O7	1
-ve	flavonol	myricetin	myr--	20.2	318.0372	318.0376	-1.23	C15H10O8	1
-ve	phenolic acid	galloyl-rhamnose	gal-rha-	4.4	318.0559	318.0587	-8.80	C12H14O10	2
-ve	stilbenoid	trans-piceide	tra-pic-	15.8	390.1314	390.1315	-0.33	C20H22O8	1
-ve	stilbenoid	cis epsilon viniferin	cis-eps-vi	20.5	454.1420	454.1416	0.88	C28H22O6	1
-ve	sugar	phospho-dihexonic acid	pho-dih-ac	4.4	482.0630	482.0673	-7.22	C13H23O17P	2
-ve	flavonol	syringetin-3-galactoside	syr-3-ga	20.4	508.1181	508.1217	-7.15	C23H24O13	1
-ve	dihydro-flavonol	dihydro-syringetin-glucoside	dih-syr-gl	16.5	510.1376	510.1372	0.78	C23H25O13	2
+ve	alkaloid	trigonelline	tri--	1.5	137.0473	137.0477	-2.92	C7H7NO2	1
+ve	aminoacids	phenyl-alanine	phe-ala-	6.0	165.0780	165.0790	-5.93	C9H11NO2	1
+ve	phenolic acid	gallic acid	gal-aci-	4.5	170.0200	170.0215	-8.82	C7H6O5	1
+ve	stilbenoid	trans-resveratrol	tra-res-	19.6	228.0801	228.0786	6.58	C14H12O3	1
+ve	aminoacids	L-glutathione	L-glu-	3.0	307.0851	307.0838	4.23	C10H17N3O6S	1
+ve	anthocyanin	cyanidin 3-glucoside	cya-3-gl	12.7	449.1085	449.1084	0.22	C21H21O11	1
+ve	anthocyanin	peonidin-3-glucoside	peo-3-gl	14.6	463.1243	463.1240	0.65	C22H23O11	1
+ve	anthocyanin	delphinidin-3-glucoside	del-3-gl	11.3	465.1043	465.1033	2.15	C21H21O12	1
+ve	anthocyanin	petunidin-3-glucoside	pet-3-gl	13.3	479.1191	479.1190	0.21	C22H23O12	1
+ve	anthocyanin	cyanidin 3-acetyl-glucoside	cya-3-ac	18.2	491.1197	491.1190	1.43	C23H23O12	1
+ve	anthocyanin	malvidin-3-glucoside	mal-3-gl	14.8	493.1345	493.1346	-0.20	C23H25O12	1
+ve	anthocyanin	delphinidin-3-acetyl-glucoside	del-3-ac	16.8	507.1144	507.1139	0.99	C23H23O13	1
+ve	anthocyanin	petunidin 3 acetyl-glucoside	pet-3-ac	18.4	521.1310	521.1295	2.88	C24H25O13	1
+ve	flavan-3-ols	procyanidin B3	pro-B3-	8.2	578.1446	578.1424	3.81	C30H26O12	1
+ve	anthocyanin	cyanidin 3-p-coumaroyl-glucoside	cya-3-p-	20.2	595.1469	595.1452	2.86	C30H27O13	1
+ve	anthocyanin	delphinidin 3-p-coumaroyl-glucoside	del-3-p-	20.0	611.1417	611.1401	2.62	C30H27O14	1
+ve	anthocyanin	cyanidin-3,5-diglucoside	cya-3,5-di	9.0	611.1625	611.1612	2.13	C27H31O16	1
+ve	anthocyanin	peonidin 3-caffeoyl-glucoside	peo-3-ca	20.2	625.1571	625.1557	2.19	C31H29O13	1
+ve	anthocyanin	delphinidin 3,5-diglucoside	del-3,5-di	8.3	627.1568	627.1561	1.12	C27H31O17	1
+ve	anthocyanin	malvidin 3-p-coumarylglucoside	mal-3-p-	20.3	639.1728	639.1714	2.19	C32H31O14	1
+ve	anthocyanin	petunidin 3,5 diglucoside	pet-3,5-di	9.4	641.1733	641.1718	2.34	C28H33O17	1
+ve	anthocyanin	malvidin-3,5-diglucoside	mal-3,5-di	11.3	655.1888	655.1874	2.14	C29H35O17	1
-ve	sugar	sucrose	suc--	1.5	342.1153	342.1162	-2.63	C12H22O11	1
-ve	sugar	cellobiose	cel--	2.2	342.1155	342.1162	-2.05	C12H22O11	1
-ve	sugar	phospho-diglucose	pho-dig-	1.4	440.0906	440.0931	-5.68	C12H25O15P	2
-ve	aroma precursor	terpenyl-pentosyl-glucoside	ter-pen-gl	21.2	448.2309	448.2308	0.22	C21H36O10	3
-ve	aroma precursor	geranic acid-rhamnosyl-glucoside	ger-aci-rh	21.1	476.2276	476.2258	3.79	C22H36O11	1
-ve	aminoacids	caftaric acid-glutathione	caf-aci-gl	7.5	617.1162	617.1163	-0.16	C23H27N3O15S	2
+ve	aroma precursor	terpendiol-glucoside	ter-glu-	20.4	332.1849	332.1835	4.29	C16H28O7	2
+ve	aroma precursor	geranic acid-arabinopyranosyl-glucoside	ger-aci-ar	21.0	462.2134	462.2101	7.14	C21H34O11	1
+ve	aroma precursor	cis-hydroxy-linalyl-arabinosyl-glucoside	cis-hyd-li	20.1	464.2268	464.2258	2.15	C21H36O11	1

Supplementary table 3: a list of the identified compounds from the grape flesh in the chapter 7.

Name	Rt (min)	Formula Theo. MW[a] ID Level	*m/z*(Relative Intensity)	Annotation	Mode	MS/MS² spectra	MAX Peak explained (ChemSpider/ PubChem)	Peak explained proposed structure
phosphohexonoic acid	1.5	C6H13O10P 276.0246 (9.62 ppm)	**275.0208** (100) 274.0160 (30) 159.0140 (15)	[M-H]-;	NEG	**MS/MS² 275:** 159.0155 [M-H-C4H5O2P] 158.0195 [M-H-C4H6O2P] 116.0064 [M-H-C6H7O5] 115.0079 [M-H-C6H6O5] 96.9661 [M-H-(GLC-H)]- 71.0167 61.0118	3	5
galloyl-rhamnose	8.0	C12H14O10 318.0559 (8.62 ppm)	**317.0487(100)** 169.0150 (25) 318.053 (13) 147.0293 (15)	[M-H]-;	NEG	**MS/MS² 317.0487:** 205.0313 [M-H-C5H4O3] 175.0525 169.0125 [M-H-RHA] 147.0338 [M-H-Gallic] 130.0240 [M-H- C7H7O6] 129.0236 [M-H-C7H8O6] 97.0107 [M-H-C11H8O5] 87.0091 [M-H-C9H10O7] 85.0319 72.9957 71.0160	2	7
dihydro-syringetin-glucoside	16.34	C23H25O13 510.1376 (0.78 ppm)	**509.1298 (100)** 510.1331913 (25) 511.1364398 (5) 347.0781877 (5)	[M-H]-;	NEG	**MS/MS² 509**:347.0772 [M-H-GLC] 346.0694 [M-H_GLC-H] 330.0794 [M-H-GLC-OH] 329.0728 [M-H-GLC-H2O] 315.0584 [M-H-GLC-CH4O] 314.0493 [M-H-GLC-CH5O] 303.0912 299.0263 261.0831 193.0198 [M-H-C14H20O8] 192.0125 [M-H-C14H21O8] 180.0504 [M-H-dihysro-syr] 167.0405 166.0349 165.0236 [M-H-C15H20O9] 153.0611 [M-H-C15H16O10] 149.0305 137.0294	4	11
phospho-diglucose	4.2	C12H25O15P 440.0931 (5.61 ppm)	**439.0833663 (100)** 440.087824 (13) 205.0353502 (20)	[M-H]-	NEG	**MS/MS² 439**: 96.9725 [Phosphate] 78.9615 [Phosphate-H2O]	0	2
terpenyl-pentosyl-glucoside	21.20	C21H36O10 448.2309 (0.22 ppm)	**447.2230(100)** 448.2276547 (22) 449.23 (3) **493.229567 (60)** 494.232167 (15)	[M-H]-; [M+FA]-	NEG	**MS/MS² 447:** 315.1850 [M-H-Pentose] 313.1965 311.1752 233.0665 [M-H-Linalool-C2H4O2] 191.0587 161.0490 [Glucose-H2O] 159.0353 149.0504 [Penstose-H] 143.0430 131.0453 [Pentose-H2O] 125.0322 119.0400 115.0356 114.0417 113.0295 101.9893 101.0293 99.0124 97.0298 95.0175 **MS/MS² 493:** 448.2326 [M- Formic acid] 447.2284 [M-Formic acid-H] 316.1868[M-Formic acid-Pentose] 315.1828 [M-H-Formic acid-Pentose]311.1740 233.0741 191.0604 [Glucose +CH2O] 179.0623 [Glucose-H] 162.0404 161.0507 [Glucose-H2O] 159.0334 149.0522 [Pentose-H] 143.0404 132.0590 131.0406 [Pentose-H] 125.0315 119.0377 114.0183 113.0291 102.0198	24	Suggested by chemspider
Caftaric acid-glutathione	7.76	C23H27N3O15S 617.1162 (-0.16 ppm)	**616.1090 (100) 617.1126982 (25) 618.1117378 (10) 619.1136888 (2.5) 484.1032156 (4)**	[M-H]-;	NEG	**MS/MS² 162**: 616.1120 485.1040 [M-H-C3H3N2O4] 484.1076 [M-H-C3H4N2O4] 467.0950 466.0965 441.1227 440.1166 273.1021 [M-H-Caftaric acid-S] 272.0938 M-H-Caftaric acid-S-H] 254.0807 [M-H-Caftaric acid-S-H2O-H] 213.0047 212.0230 211.0146 210.0911 193.0038 179.0535 169.0148 168.0268 167.0231 150.0154 149.0140 146.0516 143.0510 128.0398	6	6
terpendiol-glucoside	20.4	C16H28O7 332.1835, (4.29 ppm)	**355.1741252 (100)** 356.1763053 (17) 135.1177564 (12) 357.1824086 (4) 371.1574036 (39)	[M+Na]+ [M+K]+	POS	**MS/MS² 355**: 205.0514 [Glucose +Na(A+2)] 204.0535 [Glucose+Na(A+1)] 203.0571 [Glucose+Na] 194.1161 Terpendion +Na(A+1)] 193.1231 [Terpendiol+Na] 185.0478 [Glucose+Na-H2O] 175.1088 [Terpendiol+Na-H2O] 143.0378	3	7
digalloyl derivative	15.01	C24H16O10 464.0746 (4.96 ppm)	**463.0648 (100)** 464.0686025 (24) 465.0637019 (3) 927.1400991 (2)	[M-H]- [2M-H]-	NEG	**MS/MS² 463:** 326.0472 325.0386 [M-H-tryhydroxy-phenol] 312.0643 311.0599 [M-H-galloyl moiety] 300.0452 299.0569 295.0625 293.0488 [M-H-Gallic acid] 271.0654 253.0503 243.0691 227.0490 225.0567 170.0238 [Gallic acid (A+1)] 169.0200 [Gallic acid] 168.0109 151.0046 [Gallic acid–H2O] 145.0336 137.0296 [Gallic acid-CH4O] 126.0076 125.0292 [Trihydroxy-phenol] 124.0257	4	7
lariciresinol-rhamnoside	17.8	C26H34O10 506.2132 (3.96 ppm)	**505.2076 (100)** 506.2120384 (28) 551.2121829 (3) 569.2074749 (1)	[M-H]-; [M+FA]-	NEG	**MS/MS² 505:** 475.1643 360.1628 [M-H-Rhamnoside(A+1)] 359.1536 M-H-Rhamnoside] 345.1365 [M-H-Rhamnoside-CH2] 344.1304 342.1553 341.1439 [M-H-Rhamnoside-H2O] 329.1067 327.1286 [M-H_Rhamnoside-CH4O] 326.1220 187.0818 109.0339 101.0289 89.0264	4	12

						85.0324 73.0307 59.0169		
laricresinol-rhamnoside	18.3	C26H34O10 506.2153 (0.26 ppm)	**505.2076 (100)** 506.2120384 (28) 551.2121829 (3) 569.2074749 (1) 507.2150 (1)	[M+H]+; [M+FA]-	NEG	**MS/MS² 505:** 360.1628 [M-H-Rhamnoside(A+1)] 359.1536 M-H-Rhamnoside] 345.1365 [M-H-Rhamnoside-CH2] 344.1304 342.1553 341.1439 [M-H-Rhamnoside-H2O] 329.1067 327.1286 [M-H_Rhamnoside-CH4O] 241.0566 [C14H9O4] 187.0818 109.0339 101.0289	4	11
laricresinol-xyloside	18.9	C25H32O10 492.2037; (8.96 ppm)	**491.1965122 (100)** 492.1956684 (24) 493.1929945 (4)	[M-H]-;	NEG	**MS/MS² 491:** 477.1759 [M-H-O] 476.1733 [M-2H-O] 461.1432[M-H-O2] 359.1525[M-H-xyloside] 343.1223 [M-H-Xyloside-O] 329.1422 315.1174 314.1205 299.0951 283.1020 281.0848 270.0887	7	8
Aminoacid-derivative	16.7	C20H23NO6 373.1534 (2.41 ppm)	**374.1606 (100)** 375.1650 (21)	[M+H]+;	POS	**MS/MS²** : 223.1196 222.1154 217.0830 205.0958 204.1022 [C12H14N1O2] 202.0961[C12H12N1O2] 199.0737 187.0715 [C12H11O2] 177.0891 [C10H11N1O2] 176.0825 [C10H10N1O2] 175.0797 C10H9N1O2] 163.0541 [C9H9N1O2] 151.0645 150.0962 147.0637 124.0317 123.0490	11	
hydroxy-hexanoate-glucoside	11.4	C12H22O8 294.1309 (-2.4 ppm)	**293.1247 (100)** 294.1280 (14) 295.1300 (2) 131.0740 (3)	[M-H]-;	NEG	**MS/MS² 293**: 131.0737 [hydroxy-hexanoate-H] 113.0343 [Glucose fragment- H] 101.0276 [Glucose fragment-H] 89.0255 85.9976 [Hydroxy-hexamoate fragment-H] 85.0572 83.0188 71.0172 59.0156	6	9
galloyl-shikimic acid	7.1	C14H14O9 326.0631 (-2.2ppm)	**325.056 (100)** 326.0600 (15) 169.0150 (10)	[M-H]-;	NEG	**MS/MS² 325:** 325.0718 170.0242 [gallic acid-H(A+1)] 169.0189 [gallic acid-H] 168.0096 137.0260 [shikimic acid- H4O2-H] 126.0303 [trihydroxybenzoic acid] 125.0289 trihydroxy-benzoic acid-H] 124.0229 trihydroxy benzoic acid -2H] 123.0156 111.0463 107.0187	9	Suggested by chemspider
Galloyl-catechin A	20.5	C22H16O10 440.0733 (-2.27)	**439.0661 (100)** 440.0701 (20) 441.0720 (3) 288.0650 (10)	[M-H]-;	NEG	**MS/MS² 439:** 289.0723 [catechin-H] 288.0634 [catechin-2H] 287.0544 [Catechin- 3H] 275.0594 [Catechin –CH2] 259.0623 [Catechin-CH2-O-H]	5	Suggested by Chemspider
decyl-pentosyl-glucoside	20.8	C22H42O10 466.2774 (0.86 ppm)	**467.2702 (100)** 468.2750 (24) 303.22 (15) 161.0475 (4)	[M-H]-;	NEG	**MS/MS² 467.2702:** 448.2695 [M-H-H20 (A+1)] 447.2636 [M-H-H2O] 405.2608 [M-H-CH4O2] 303.2200 [M-H-Glucose-2H] 285.2112 [M-H-Glucose-H2O] 243.1977 161.0525 [Glucose-H-H2O] 159.0325 113.0282 101.0276	16	Suggested by Chemspider
mascaroside-like	18.9	C26H36O11 524.2249 (-1.71 ppm)	**523.2177 (100)** 524.2200 (26) 361.170567 (16) 343.157367 (8)	[M-H]-,	NEG	**MS/MS² 523.2177**:362.1693 [M-H-Glucose (A+1) 361.1716 [M-H-Glucose] 347.1563 [M-H-CH2-Glucose] 346.1458 M-2H-CH2-Glucose] 343.1573 [M-H-Glucose-H2O] 313.1487 165.0570 122.0492	5	9
vanilliyl-glucoside	11.4	C14H20O8 316.1158 (-4 ppm)	**317.1220 (100)** 318.1246583 (16) 339.1060 (20)	[M+H]+; M+Na	POS	**MS/MS² 317**:205.0405 204.0688 203.0580 (Glucose+Na] 187.0155 186.0455 185.0455 [Glucose+Na-H2O] 165.0574 157.0332 156.0775 [Vanillyl-alcohol-H (A+1)] 155.0712 [Vanyllil-alcohol-H] 151.0431 149.0661 [Vanyllil-alcohol-O+CH2-H] 145.0437 143.0357 127.0444 123.0470 109.0360 99.0520 97.0342 91.0136	10	14
hexyl-rhamnosyl-glucoside	20.0	C18H34O9 394.2203 (-0.35 ppm)	**417.2093619 (100)** 418.21227462 (20) 203.0586 (11)	[M+Na]+	POS	**MS/MS 417:** 297.1496 256.1774 255.1542 [M+Na-Glucose] 239.1199 238.1434 [M+Na-Glucose-H2O (A+1)] 237.1446 [M+Na-Glucose-H2O] 204.0571 203.0579 [Glucose+Na] 185.0475 [Glucose+Na-H2O] 143.0417	5	Suggested by Chemspider
dihydrokaempferol-glucoside	12.0	C21H22O11 450.1162 (3.11 ppm)	**473.1068 (100)** 474.1108477 (23) 489.0878 (14)	M+Na M+K	POS	**MS/MS 473:**338.1057 337.0925 323.0742 314.0707 313.0886 312.0668 [M+Na-Glucose (A+1)] 311.0560 [Dihydro-kaempferol+Na] 293.0517 [M+Na-Glucose-H2O] 265.0542 185.0472 [Glucose-H20+Na] 97.0299 85.0335	3	5
Oxalyl-benzoic acid	11.4	C9H6O5 194.0213 (1.15 ppm)	**193.0140 (100)** 194.017685 (10)	M-H	NEG	**MSMS 193:** 175.0057 [M-H-H2O] 165.0229 163.9817 147.0123 [M-H-CH2O] 138.0759 137.0287 [hydroxy-benzoic acid] 135.9881 135.0160 134.0417 122.0441 121.0256 [benzoic acid] 119.0140 111.0110 110.0189 109.0318 107.0147 [benzoic acid-O+2H] 93.0382 92.0184 91.0211 89.0295	9	Suggested by Chemspider
dihydro-myricetin	6.8	C15H12O8	**319.0454 (100)**	M-H	NEG	**MS/MS 319:** 215.0397 [C12H7O4] 194.0318 [C9H8O5] 193.0180 [C9H7O5]	19	

		320.0526 (-1.93 ppm)	320.0490746 (16) 193.0180 (15)			192.0440 191.0371 190.0242 187.0402 185.0423 167.0372 165.0238 163.0273 159.0437 153.0176 151.0041 147.0149 145.0378 139.0445 137.0271 **[**dihydroxy-benzoic moiety**]** 135.0307 126.0243 125.0275 123.0328 121.0332 **[**benzoic moiety**]** 119.0245 110.0310 109.0312 107.0228		Suggested by KEGG
galloyl-syringic acid	8.6	C16H14O9 350.0633 (1.14 ppm)	**349.0561 (100)** 350.05087435 (18) 169.0150 (12) 197.0350 (9)	M-H	NEG	**MS/MS 349:** 274.0334 273.0462 231.0352 227.0379 205.0528 203.0323 199.0461 192.0065 191.0016 189.0483 187.0422 177.0295 **[**syringic acid-H2O**]]** 175.0409 166.0250 165.0208 164.0143 163.0096 159.0519 153.0218 152.0075 151.0089 145.0428 137.0259 135.0180 131.0513 125.0256 124.0181 121.0364 109.0318 107.0239	22	Suggested by Chemspider
galloyl-ethyl-gallate	14	C16H14O9 350.0633 (-1.43 ppm)	**349.0550 (100)** 350.0550	M-H	NEG	**MS/MS 349:** 197.0487 166.0276 165.0216 151.0073 **[**gallic acid-H2O-H**]** 138.0304 137.0267 **[**hydroxy-benzoic acid**]** 123.0107 121.0319 **[**benzoic acid**]** 109.0332 97.0307 95.0179 93.0338 83.0155	19	Suggested by Chemspider
quercetin-rhamnoside	16.9	C21H20O13 448.0985 (-4.6 ppm)	**447.0878 (100)** 448.09237465 (23) 315.054535 (10) 301.0350 (20) 300.02874356 (4)		NEG	**MSMS 447:** 341.0685 316.0704 315.054 **[**M-H-dehidro-rhamnose**]** 314.0471 301.0299 **[**Quercetin-H**]** 300.0309 **[**Quercetin-2H**]** 299.0235 **[**Quercetin-3H**]** 297.0742 288.0624 287.0638 286.0515 285.0493 **[**M-H-Rhamnose-H2O**]** 284.0398 273.0217 271.0374 269.0602 243.0345 241.067 229.047 227.0652 217.0194 193.022 189.0311 177.0191 175.0124 165.0269 163.0448 **[**Rhamnose-H**]** 161.0356 **[**Rhamnose-3H**]** 152.0114 151.0065 **[**Gallic acid-H2O-H**]** 136.0451 135.0462 125.0249 109.0342	51	Suggested by Chemspider
ellagic acid-glucoside	15.9	C20H16O13 464.0586 (-1.06 ppm)	**463.0510 (100)** 464.05689364 (22) 300.997653 (12) 299.995654 (4)	M-H	NEG	**MS/MS 463:**437.1108 **[**M-CO-H**]** 313.0951 303.0222 302.0057 **[**Ellagic acid**]** 300.9983 **[**Ellgaic acid-H**]** 299.9916 **[**Ellagic acid-2H**]** 299.0925 285.0426 275.0595 259.0651 152.0171 151.0233 125.0266 89.0257	6	Suggested by Chemspider
galloyl-ellagic acid	14	C21H10O13 470.0122 (0.13 ppm)	**469.0050 (100)** 470.0087654 (21) **425.0125 (500)** 426.016758 (100) 300.998756 (20)	M-H M-CO2-H	NEG	**MS/MS 469:** 451.1294 425.0125 **[**M-H-CO2**]** 303.0112 302.0080 **[**Ellagic acid**]** 301.0015 **[**Ellagic acid-H**]** 299.9934 **[**Ellagic acid-2H**]** 298.9927 285.0057 **[**Ellagic acid-H-CH2**]** 283.9969 282.9954 273.0014 271.9992 270.9910 257.0063 245.0154 243.0300 229.0212 219.0294 216.0098 201.0220 200.0146 161.0298 125.0295 **[**Benzoic acid**]** **MS/MS 425:** 302.9927 301.9973 **[**Ellagic acid**]** 301.0002 **[**Ellagic acid-H**]** 299.9947 **[**Ellagic acid-2H**]** 283.9972 **[**Ellagic acid-H2O**]** 282.9940 272.9931 271.9967 245.0138 244.0061 243.0097 229.0238 227.0151 201.0567 200.0225 187.0664 173.0143	10	Suggested by Chemspider
dihydro-syringetin-glucoside	16.5	C30H22O8 510.1354 (7.71 ppm)	**509.1282 (100)** 510.1299 (29) 511.1310 (3) 347.08756 (20)	M-H	NEG	**MS/MS 509** 355.0693 348.0843 **[**dihydrosyringetin**]** 347.0771 (**[**dihydro-syringetin-H**]** 346.0728 331.0714 330.0712 **[**dihydro-syringetin-H2O**]** 329.0698 **[**dihydrosyringetin-H2O-H**]** 319.0854 315.0520 dihydrosyringetin-CH4O**]** 314.0435 303.0898 299.0224 261.0791 220.0393 205.0171 193.0161 192.0078 180.0456 **[**Glucose???**]** 167.0360 166.0272 165.0208 164.0142 153.0580 150.0302 149.0258 138.0324 137.0264 **[**Hydroxy-benzoic acid**]** 125.0259 **[**glucose fragment**]**	6	12
HHDP-galloyl-glucose	4.5/5.3 /8.1	C27H22O18 634.0802 (-0.65 ppm)	**633.073 (100)** 634.0776 (29) 300.99764 (10) 463.0654 (3)	M-H	NEG	**MS/MS 633**:481.0687 **[**M-H-Galloyl**]** 463.0608 **[**M-H-Gallic acid**]** 421.0453 419.0682 331.0718 303.0038 302.0079 301.0023 **[**Ellagic acid**]** 300.0663 **[**Ellagic acid-H**]** 277.0198 276.0275 275.0229 273.0112 257.0088 251.0579 250.0462 249.0435 231.0337 211.0278 169.0164 **[**Gallic acid-H**]** 125.0280 **[**Glucose gfragment**]**	13	Suggested by Chemspider
punicalin-like	11.5	C34H22O22 782.0623 (2.59 ppm)	**781.0551 (100)** 782.0586735 (36) 783.0612233 (6) 300.9975634 (10) 463.0543 (5)	M-H	NEG	**MS/MS 781:**765.0464 **[**M-H-O**]** 764.0444 **[**M-2H-O**]** 763.0395 **[**M-H-H2O**]** 754.0591 753.0580 **[**M-CO-H**]** 748.0308 747.0374 746.0352 745.029 737.0562 M-CO2**]** 736.0515 735.0482 720.0582 719.0544 708.0523 707.0560 701.0424 691.0585 479.0495 M-Ellagic acid**]** 463.0552 **[**M-Ellagic acid-O**]** 462.0464 461.0379 M-Ellagic acid-H2O**]** 451.0526 445.0413 443.0323 **[**M-HHDP**]** 417.0481 316.9964 303.0076 302.0067 **[**Ellagic acid**]** 301.0020 **[**Ellagic acid-H**]** 300.0015 298.9863 291.0196 289.0012 275.0233 274.0163 273.0083 **[**Ellagic acid fragment**]** 249.0442 247.0278	12	Suggested by Chemspider

						245.0166 22		
HHDP-digalloyl-glucose	10.5	C34H26O22 786.0962 (5.89 ppm)	**785.089 (100)** 786.093864 (35) 787.010273 (5) 300.989765 (10)	M-H	NEG	**MS/MS 785:** 766.0676 **[**M-H2O-H**]** 625.1403 **[**M-Gallic acid**]** 589.0814 483.0812 **[**Ellagic acid-glucoside+H2O**]** 463.0885 Ellagic acid-Glucoside-H**]** 419.0655 302.0099 **[**Ellagic acid**]** 301.0025**[**Ellagic acid-H**]**275.0262 **[**Ellagic acid fragment**]** 249.0452	4	Suggested by Chemspider
Heterenoside	13.3	C36H38O20 790.1975 (2.50 ppm)	**789.1903 (100)** 790.19507365 (40) 463.0878 (17) 301.0334 (11)	M-H	NEG	**MS/MS 789**:663.1583 643.1424 **[**M-H-Rhamnose**]** 629.1439 628.1417 627.1370 **[**M-H-Glucose**]** 611.1267 610.1361 609.1273 M-H-Glucose-H2O**]** 503.1160 502.1103 501.1051 483.0960 482.1046 481.1018 **[**M-H-Rhamnosyl-glucoside**]** 477.1270 **[**Quercetin-glucoside+CH2**]** 476.1307 475.1255 464.0935 463.0915 **[**Quercetin-glucoside**]** 355.0696 329.0889 320.0546 319.0474 307.0843 302.0457 **[**Quercetin**]** 301.0379 **[**Quercetin-H**]** 300.0346 284.0268 283.0292 **[** Quercetin- H2O**]** 265.0738 193.0164 192.0098 176.0021 175.0054 167.0376 149.0259 125.0272	15	Suggested by Chemspider
Hydroxy-benzodioxine-carboxylic acid	9.8	C9H6O5 194.0235 (9.97 ppm)	**195.0306581 (100)** 196.033524 (10) 121.0333454 (3)	M+H	POS	**MS/MS 195:** 198.2060 177.0220 **[**M+H-H2O**]** 167.0644 **[**M-CO**]** 150.0410 **[**M-CO2**]** 149.0274 **[**M-CO2+H**]** 140.0050 139.0414 **[**M-Hydroxy-benzoic acid**]** 133.0361 123.0333 122.0076 121.0340 **[**M-benzoic acid-H**]** 111.0570 107.0157 105.9884 105.0396 102.9831 95.0558	14	Suggested by Chemspider
dihydro-myricetin	10	C15H12O8 320.0542 (2.96 ppm)	**321.0613645** 322.06554 (17) 139.034084 (4)	M+H	POS	**MS/MS 321:** 229.0527 220.0361 219.0332 196.0342 195.0336 191.0387 177.0228 167.0376 163.0437 154.9987 154.0153 153.0225 151.0383 149.0275 143.0384 140.0371 139.0443 **[**hydroxy-benzoic acid**]** 133.0335 127.0394 126.0162 125.0291 123.0429 121.0340 111.0349 107.0173 105.0388 97.0336 93.0397	16	Suggested by Chemspider
dihydro-laricitrin	9.32	C16H14O8 334.0695 (1.89 ppm)	**335.0755 (100)** 336.079136 (16)	M+H	POS	**MS/MS 335:** 233.0493 209.0496 194.0263 181.0545 177.0370 168.0376 167.0387 153.0580 150.0357 149.0293 140.0386 139.0444 **[**Hydroxy-benozic acid**]** 138.0403 111.0406 65.0440	15	17
Dihydro-syringetin	11.8	348.0860 (4.33 ppm)	**349.0932234 (100)** 350.097765 (18) 247.045362 (13)	M+H		**MS/MS 349:** 247.0654 223.0659 208.0401 195.0501 190.0322 183.0279 182.0587 181.0542 177.0271 163.0390 162.0374 154.0573 153.0603 149.0281 140.0401 139.0441 138.0376 125.0581 123.0320 121.0349 111.0455 110.0380 108.0304 107.0544	16	16
ellagic acid-arabinoside	19.2	C19H14O12 434.0499 (3.24 ppm)	**435.0571335 (100)** 436.0610 (20) 302.01112 (10)	M+H	POS	**MS/MS 435:** 331.0601 305.0182 304.0201 **[**Ellagic acid +H (A+1)**]** 303.0198 **[**Ellagic acid +H**]** 285.0123 **[**Ellagic acid-H2O+H**]** 275.0301 257.0262 **[**Ellagic acid-CO2+H**]**	4	Suggested by Chemspider
cyanidin-3-arabinoside	15.6	C20H19O10 419.0973 (-1.26 ppm)	**419.0973 (100)** 420.10342 (20) 287.0565 (14)	M+H	POS	**MS/MS 419:** 373.0276 290.0519 289.0577 288.0650 **[**Cyanidin A+1)**]** 287.0605 **[**Cyanidin**]** 147.0514 **[**Rhamnose-H2O+H**]**	2	Suggested by Chemspider
delphinidin-3-arabinoside	13.9	C20H19O11 435.0944 (3.72 ppm)	**435.0944 (100)** 436.0978645 (20) 303.051123 (16)	M+H	POS	**MS/MS 435:** 303.0530 **[**Delphinidin**]** 304.0627 **[**Delphinidin +H**]]** 305.0520	3	Suggested by Chemspider
malvidin-glycosyl-glucoside	8.8	C29H31O17 651.1562 (0.06 ppm)	**651.1562 (100)** **652**.1601(31) 653.1634 (4) 331.0876 (15)	M+H	POS	**MS/MS 651:**635.1502 **[**M-O**]** 634.1542 **[**M-O-H**]** 633.1481 M-H2O**]** 491.1102 **[**Malvidin-glucoside-2H**]** 490.1092 489.1054 474.0948 473.100 472.1013 471.0943 331.0855 **[**Malvidin**]** 328.0587 327.0519 310.0512 309.0421Malvidin-H2O**]**	9	Suggested by Chemspider
peonidin-diglucoside-p-coumaroyl	19	C37H39O18 771.2170 (4.33 ppm)	**771.2170 (100)** 772.2213 (40) 773.2235 (4) 301.0745 (11)	M+H	POS	**MS/MS 771** 611.1656 610.1657 609.1630 **[**M-Glucose**]** 463.1271 **[**P-coumaroyl-peonidin**]** 303.0682 302.0848 **[**Peonidin+H**]** 301.0751 **[**Peonidin**]**	5	6
catechin-	12.2	C21H24O10	**435.1289079 (100)**	M-H	NEG	**MS/MS 435:** 313.0890 297.0961 289.0713 **[**Catechin-H**]** 271.0658 Catechin-H-	9	

rhamnoside		436.1361 (1.92 ppm)	436.1329373 (24) 437.1334 (2) 289.0720 (19)			H2O] 245.0837 [Catechin-H-CO2] 227.0715 Catechin-H-CH4O2] 221.0812 205.0489 203.0765 187.0414 185.0568 179.0324 167.0398 166.0307 165.0293 164.0126 163.0401 [Rhamnose-H] 161.0735 159.0434 153.0129 152.0343 151.0412 150.0339 149.0276 146.0351 145.0331 [Rhamnose-H-H2O] 139.0179 138.0242 137.0260 [hydroxy-benozic acid-H] 126.0244 125.0261 123.0437 122.0488 121.0351 109.0334		16
terpendiol-glucoside	17.6	C16H28O7 332.1847 (2.4 ppm)	**355.1735 (100)** 356.1776 (18)	M+Na	POS	**MS/MS 355:** 205.0313 204.0676 [Glucose+Na+H] 203.0574 [Glucose+Na] 83.0864 71.0528	0	3
hydroxy decanoic acid pentosyl-glucoside	12.3	C21H38O12 482.2375 (2.44 ppm)	**505.2267052 (100)** 506.2301 (23) **373.1856825 (80)** 374.1888 (16) 521.2034 (11)	M+Na M+Na-Pentosyl M+K	POS	**MS/MS 505:** 373.1861 [M+Na-Pentosyl] 336.1066 335.0998 334.0900 333.0823 **MS/MS 373:** 211.1455 M+Na-Glucose] 204.0597 203.0568 [Glucose+Na] 201.0430 195.1261 [Decanoic acid+Na] 193.1286	1	4
Terpendiol-rhamnosyl-glucoside*	19.5	C22H38O11 478.2414 (0.13 ppm)	**523.2400 (100)** 524.244762 (24) 525.2477645 (3) 477.2351 (43)	M+FA- M-H	NEG	**MS/MS 523:** 479.2228 478.2369 [M-H (A+1)] 477.2343 [M-H] 331.1855 [M-H-Rhamnose] 265.0771 247.0849 [Glucose+Rhamnose fragment] 206.1077 205.0695 185.1301 [Terpendiol+H2O-H] 179.0371 [Glucose-H] 164.0797 163.068 [Rhamnose-H] 161.0429 [Glucose-H-H2O] 149.0659 [Rhamnose-H-H2O] 145.0542 143.0243 131.036 119.0416 113.0245 103.0419 101.0325	4	12
Terpendiol-pentosyl-glucoside*	16.8 17.9 19.9 16.2 21.0	C21H36O11 463.2257 (1.5 ppm)	**509.2255 (100)** 510.2287654 (22) 463.218676 (46) **463.2200 (100)** 464.2245 (22) 331.1754 (5)	M+FA- M-H M-H	NEG	**MS/MS 509:** 465.183 464.2178 463.2182 [M-H] 332.1707 [M-H-Pentose (A+1)] 331.1744 [M-H-Pentose] 191.0867 179.0393 [Glucose-H] 161.0576 [Glucose-H-H2O] 149.056 [Pentose-H] 101.0467 **MS/MS 463:** 333.1836 332.1815 331.1783 [M-H-Pentose] 286.0136 233.0507 [Glucose-Pentose fragment] 161.9618 161.0508 [Glucose-H-H2O] 159.0265 149.0419 [Pentose-H] 143.0351 Glucose-H2O-H2O-H] 131.0437 Pentose-H2O-H] 125.0253	5	11
Hydroxy-citronellol-pentosyl-glucoside	19.4 20.2	C21H40O11 466.2414 (-2.3 ppm)	**489.2300 (100)** 490.2325 (24) 491.2400 (2) 357.18 (17) **511.2473 (100)** 512.2534 (22) 465.2349 (27)	M+Na M+FA	POS	**MS/MS 489:** 413.0481 373.1477 358.1807 357.1808 [M+Na-Pentose] 356.1729 355.1395 345.0293 340.9822 337.0691 336.1034 335.0938 [Glucose+Pentose+Na] 334.0291 333.0687 329.0971 325.0584 304.0634 303.0487 276.0602 275.0596 203.0573 [Glucose+Na] 201.074 185.0307 [Glucose+Na-H2O] **MS/MS 511:**469.1476 468.257 467.2536 465.2313 [M-H] 336.1748 335.201 333.1883 [M-H-Pentose] 289.0893 233.0618 [Glucose-Pentose fragment] 161.0433 [Glucose-H2O-H] 149.0462 [Pentose-H] 113.0256 101.0245	3	8
oxo-alpha-ionol-glucoside	19.9	C19H30O7 370.1991 (2.1 ppm)	**415.1987 (100)** 416.2011 (22) 369.1889 (20)	M+FA M-H	NEG	**MS/MS 415:**401.2591 399.0708 398.1806 369.1867 [M-H] 356.1707 318.9537 312.175 299.9476 295.1591 [Glucose+C9H8O] 294.0438 291.0456 288.1522 287.1559 [Glucose+C8H8] 285.8146 284.789 229.0768 228.2395 227.1234 225.9824 225.137 223.1342 [Oxo-alphaionol+O-H] 207.1432 [Oxo-alpha-ionol-H] 175.0081 173.0559 170.0411 161.0415 [Glucose-H-H2O] 146.0767 144.1754 137.0207 129.0435 128.0341	5	8
Homovanyllil alcohol-pentosyl-glucoside	12.41	C20H32O12 462.1737 (6.8 ppm)	**461.1598 (100)** 462.1645 (22) 329.117635 (11)	M-H	NEG	**MS/MS 461:** 338.9826 329.1046 M-H-Pentose] 313.1206 [homovanillyil-alcohol-Glucoside-O-H] 285.0401 271.0233 257.0353 256.027 241.0523 217.0406 201.0254 199.0404 198.2018 197.1797 193.0133 192.0023 179.0599 [Homovanillyil-alcohol+C-H] 175.046 167.0649 [Homovanillyil-alcohol-H] 165.0347 [Homovanillyil-alcohol-H-2H] 149.0228 [Pentose-H] 137.0259 [Homovanillyil-alcohol-CH2O-H] 125.0622	5	6
Hotrienol-Pentosyl-glucoside	20.6	C21H34O10 446.2151 (-0.45 ppm)	**445.2087 (100)** 446.2136	M-H	NEG	**MS/MS 445:**285.1362 284.139 283.1563 [M-H-Glucose] 265.1291 [M-H-Glucose-H2O] 239.1744 222.0985 221.1524 210.1087 209.115 [Hotrienol+Glucose fragment]	2	5
Vomifoliol-pentosyl-glucoside		C24H38O12 518.2363 (1.32	**517.2311 (50)** 518.2346 (13)	M-H	NEG	**MS/MS 517:** 359.1458 205.1237 [vomifoliol-H2O-H] 191.0522 [Glucose-H2O+CH2O] 161.0448[Glucose-H2O-H] 149.0418 [Pentose-H]	5	

		ppm)	205.1411 (7) 563.2368 (22) **431.192 (100)** 432.19623 (20)	M+FA M-Pentose+FA		101.0238 **MS/MS 431:** 385.1742 **[**M-Pentose-H**]** 363.1583 345.1869 223.1298 **[**Vomifoliol-H**]** 217.0282 205.1242 Vomifoliol-H2O**]** 179.0578 **[**Glucose-H**]** 161.0433 **[**Glucose-H2O-H**]** 153.096 **[**Vomifoliol-fragment**]** 152.0783 151.0945 Vomifoliol-Fragment-2H**]** 119.0348		12
Terpentriol-rhamnosyl-glucoside	20.1 19.7	C22H38O12 494.2363 (3.24 ppm)	**539.2339 (100)** 540.2387 (24) 493.2301 (13)	M+FA	NEG	**MS/MS 539:**497.2139 496.2522 495.2361 494.2461 **[**M-**]** 493.2288 **[**M-H**]** 40.069 379.132 350.1853 349.183 **[**M-H-Rham nosyl+2H**]** 331.1492 **[**M-H-Glucose**]** 325.0518 307.0744 **[**Rhamnosyl-glucoside-H**]** 266.0093 265.0947 **[**Rhamnosyl-glucoside fragment**]** 247.0843 **[**Rhamnosyl-glucoside fragment**]** 235.0804 205.0823 187.1315 185.117 **[**Terpentriol-H**]** 163.0651 **[**Rhamnosyl-H**]** 161.0464 **[**Glucose-H-H2O**]** 119.0321	5	14
Terpentriol-pentosyl-glucoside	17.29	C21H36O12 480.2206 (1.12 ppm)	**479.2116** 480.2156 (23) 185.1160 (9)	M-H	NEG	**MS/MS 479:**316.0214 293.1028 **[**Pentosyl-glucoside-H**]** 289.1493 265.0949 260.9919 235.0092 233.0637 **[**Pentosyl-glucoside fragment**]** 205.0721 **[**Pentosyl glucoside fragment –CO**]** 191.0515 185.1162 **[**Terpentriol-H**]** 183.0969 Terpentriol-H-2H**]** 179.0624 **[**Glucoside-H**]** 161.0570 **[**Glucose-H-H2O**]** 149.0360 **[**Pentose-H**]** 143.0355 131.0327 **[**Pentose-H-H2O**]** 125.0230 121.0366	6	12
Hydroxy-Geranic acid-Rhamnosyl-glucoside	17.7	C22H36O12 492.2207	**537.2234 (100)** 538.2277 (22) 491.2135 (20)	M+FA M-H	NEG	**MS/MS 537:** 491.2155 **[**M-H**]** 308.1064 307.1035 **[**Rhamnosyl-glucoside-H**]** 247.0844**[**Rhamnosyl-glucoside fragment**]** 184.1078 **[**Hydroxy-geranic acid**]** 183.1011 **[**Hydroxy-geranic acid-H**]** 163.0522 **[**Rhamnose-H**]** 161.0476 **[**Glucose-H-H2O**]** 145.0402 **[**Rhamnose-H-H2O**]**	2	9

Supplementary table 4: MS and MS/MS spectra of the unknown biomarkers studied in this thesis. The spectra have been analyzed through MetFrag (http://msbi.ipb-halle.de/MetFrag/) using ChemSpider, PubChem or Kegg as reference databases. Every identification have been confirmed through further manual curation.